Galaxy

for high-throughput sequence data analysis

The only four things you need to remember:

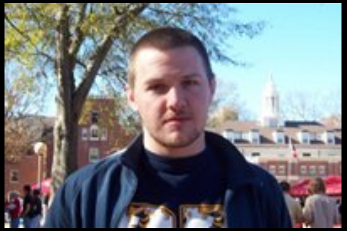
http://usegalaxy.org http://usegalaxy.org/**galaxy101**

http://getgalaxy.org http://usegalaxy.org/**cloud**

The Galaxy Team

Enis Afgan

Guru Ananda



Dannon Baker

Dan Blankenberg

Ramkrishna Chakrabarty

Nate Coraor

Jeremy Goecks

Greg von Kuster

Kanwei Li

Kelly Vincent

Supported by the National Human Genome Research Institute (HG005542, HG004909, HG005133), the National Science Foundation (DBI-0850103), Penn State University, Emory University, and the Pennsylvania Department of Public Health

A crisis in genomics research: **reproducibility**

Microarray Experiment Reproducibility

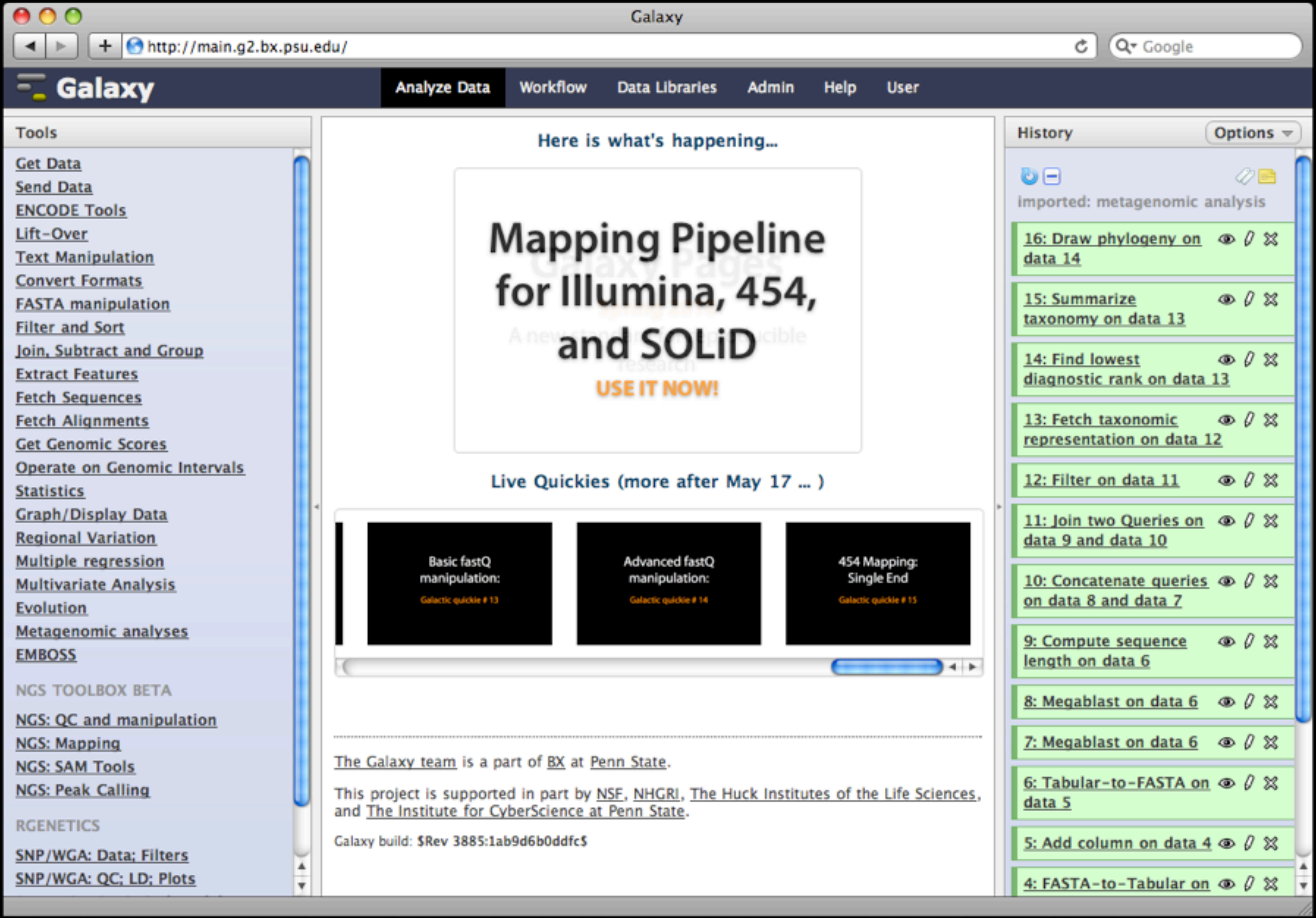
- 18 Nat. Genetics microarray gene expression experiments
- Less than 50% reproducible
- Problems
  - missing data (38%)
  - missing software, hardware details (50%)
  - missing method, processing details (66%)

Ioannidis, J.P.A. et al. Repeatability of published microarray gene expression analyses. Nat Genet 41, 149-155 (2009)

NGS Re-sequencing Experiment Reproducibility

- 14 re-sequencing experiments in Nat. Genetics, Nature, and Science (2010)
- 0% reproducible?
- Problems
  - limited access to primary data (50%)
  - some or all tools unavailable (50%)
  - settings & versions not provided (100%)

Galaxy: accessible analysis system



Here is what's happening...

Mapping Pipeline for Illumina, 454, and SOLiD

USE IT NOW!

Live Quickies (more after May 17 ...)

Basic fastQ manipulation: Galactic quickie # 13

Advanced fastQ manipulation: Galactic quickie # 14

454 Mapping: Single End Galactic quickie # 15

The Galaxy team is a part of BX at Penn State.

This project is supported in part by NSF, NHGRI, The Huck Institutes of the Life Sciences, and The Institute for CyberScience at Penn State.

Galaxy build: $Rev 3885:1ab9d6b0ddfc$

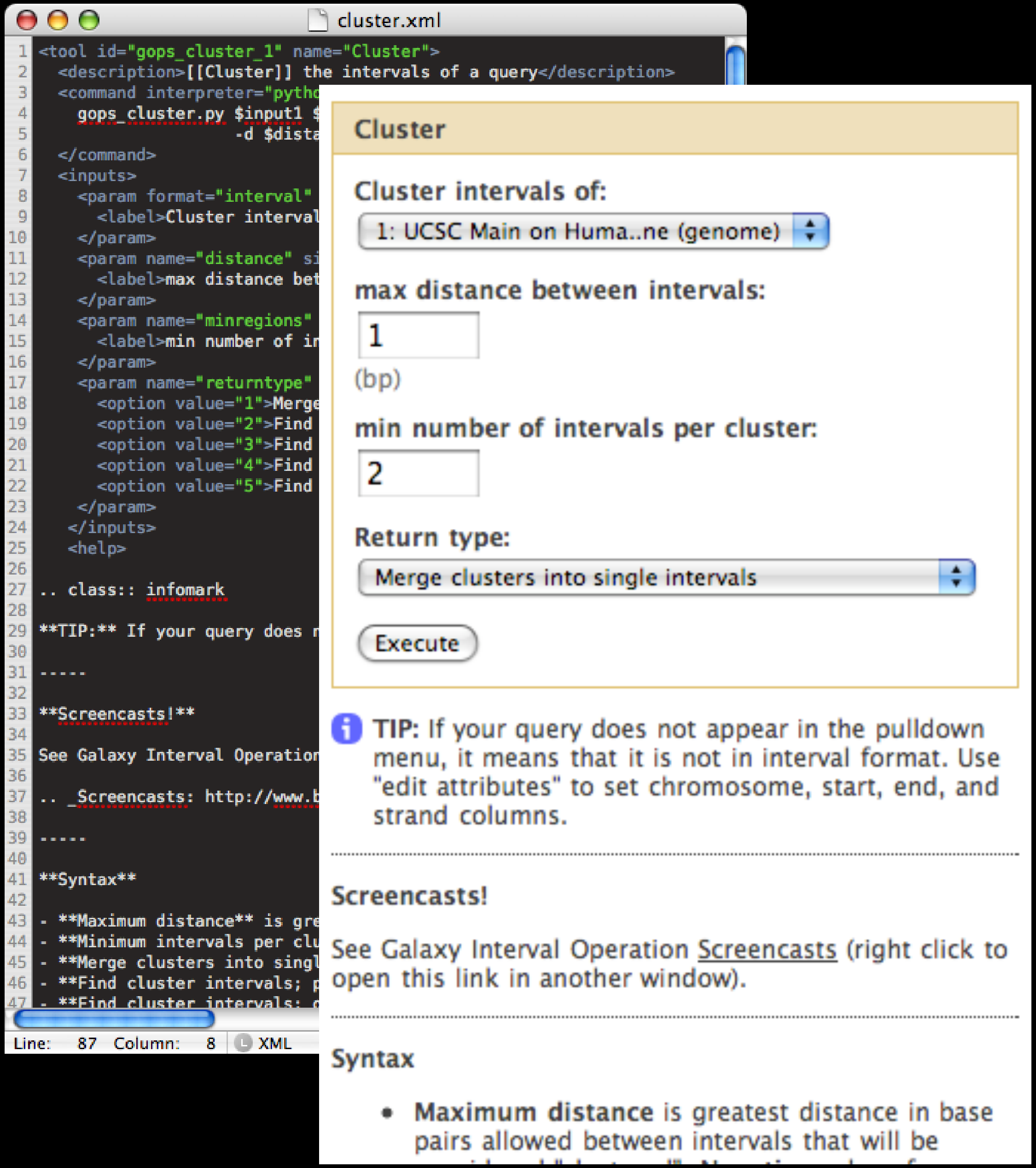
imported: metagenomic analysis

- 16: Draw phylogeny on data 14
- 15: Summarize taxonomy on data 13
- 14: Find lowest diagnostic rank on data 13
- 13: Fetch taxonomic representation on data 12
- 12: Filter on data 11
- 11: Join two Queries on data 9 and data 10
- 10: Concatenate queries on data 8 and data 7
- 9: Compute sequence length on data 6
- 8: Megablast on data 6
- 7: Megablast on data 6
- 6: Tabular-to-FASTA on data 5
- 5: Add column on data 4
- 4: FASTA-to-Tabular on

What is Galaxy?

- **A free (for everyone) web service** integrating a wealth of tools, compute resources, terabytes of reference data and permanent storage
- **Open source software** that makes integrating your own tools and data and customizing for your own site simple

Integrating existing tools into a uniform framework



cluster.xml

```
<tool id="gops_cluster_1" name="Cluster">
  <description>[[Cluster]] the intervals of a query</description>
  <command interpreter="pytho
    gops_cluster.py $input1 $
                    -d $dista
  </command>
  <inputs>
    <param format="interval"
      <label>Cluster interval
    </param>
    <param name="distance" si
      <label>max distance bet
    </param>
    <param name="minregions"
      <label>min number of ir
    </param>
    <param name="returntype"
      <option value="1">Merge
      <option value="2">Find
      <option value="3">Find
      <option value="4">Find
      <option value="5">Find
    </param>
   </inputs>
   <help>

.. class:: infomark

**TIP:** If your query does r

-----

**Screencasts!**

See Galaxy Interval Operation

.. _Screencasts: http://www.b

-----

**Syntax**

- **Maximum distance** is gre
- **Minimum intervals per clu
- **Merge clusters into singl
- **Find cluster intervals; p
- **Find cluster intervals; c
```

Line: 87 Column: 8 XML

Cluster

Cluster intervals of: 1: UCSC Main on Huma..ne (genome)

max distance between intervals: 1

(bp)

min number of intervals per cluster: 2

Return type: Merge clusters into single intervals

Execute

TIP: If your query does not appear in the pulldown menu, it means that it is not in interval format. Use "edit attributes" to set chromosome, start, end, and strand columns.

Screencasts!

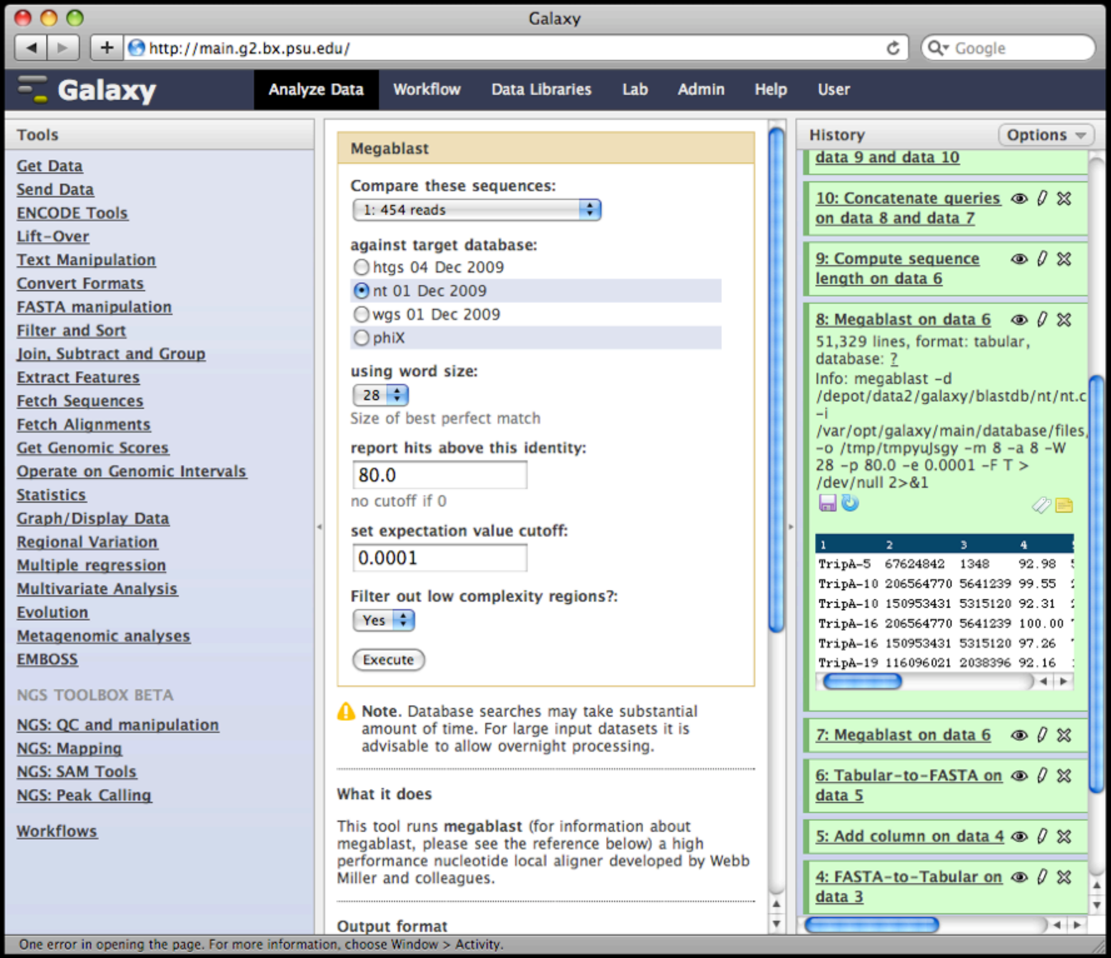
See Galaxy Interval Operation Screencasts (right click to open this link in another window).

Syntax

- **Maximum distance** is greatest distance in base pairs allowed between intervals that will be

- Defined in terms of an abstract interface (inputs and outputs)
  - In practice, mostly command line tools, a declarative XML description of the interface, how to generate a command line
- Designed to be as easy as possible for tool authors, while still allowing rigorous reasoning

Galaxy analysis interface



Megablast

Compare these sequences: 1: 454 reads

against target database:
○ htgs 04 Dec 2009
● nt 01 Dec 2009
○ wgs 01 Dec 2009
○ phiX

using word size: 28

Size of best perfect match

report hits above this identity: 80.0

no cutoff if 0

set expectation value cutoff: 0.0001

Filter out low complexity regions?: Yes

Note. Database searches may take substantial amount of time. For large input datasets it is advisable to allow overnight processing.

What it does

This tool runs **megablast** (for information about megablast, please see the reference below) a high performance nucleotide local aligner developed by Webb Miller and colleagues.

Output format

data 9 and data 10

10: Concatenate queries on data 8 and data 7

9: Compute sequence length on data 6

8: Megablast on data 6

51,329 lines, format: tabular, database: ?

Info: megablast -d /depot/data2/galaxy/blastdb/nt/nt.c -i /var/opt/galaxy/main/database/files, -o /tmp/tmpyuJsgy -m 8 -a 8 -W 28 -p 80.0 -e 0.0001 -F T > /dev/null 2>&1

1	2	3	4
TripA-5	67624842	1348	92.98
TripA-10	206564770	5641239	99.55
TripA-10	150953431	5315120	92.31
TripA-16	206564770	5641239	100.00
TripA-16	150953431	5315120	97.26
TripA-19	116096021	2038396	92.16

7: Megablast on data 6

6: Tabular-to-FASTA on data 5

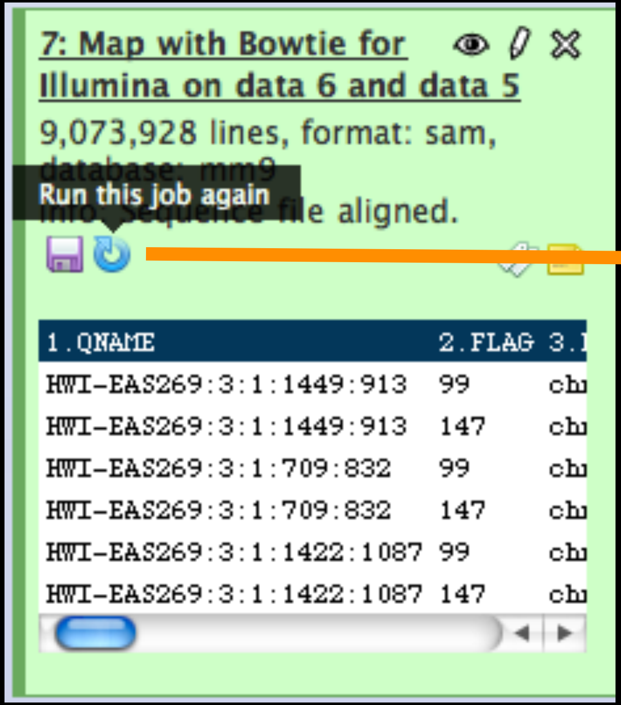
5: Add column on data 4

4: FASTA-to-Tabular on data 3

One error in opening the page. For more information, choose Window > Activity.

- Consistent tool user interfaces automatically generated
- History system facilitates and tracks multistep analyses

Automatically tracks every step of every analysis

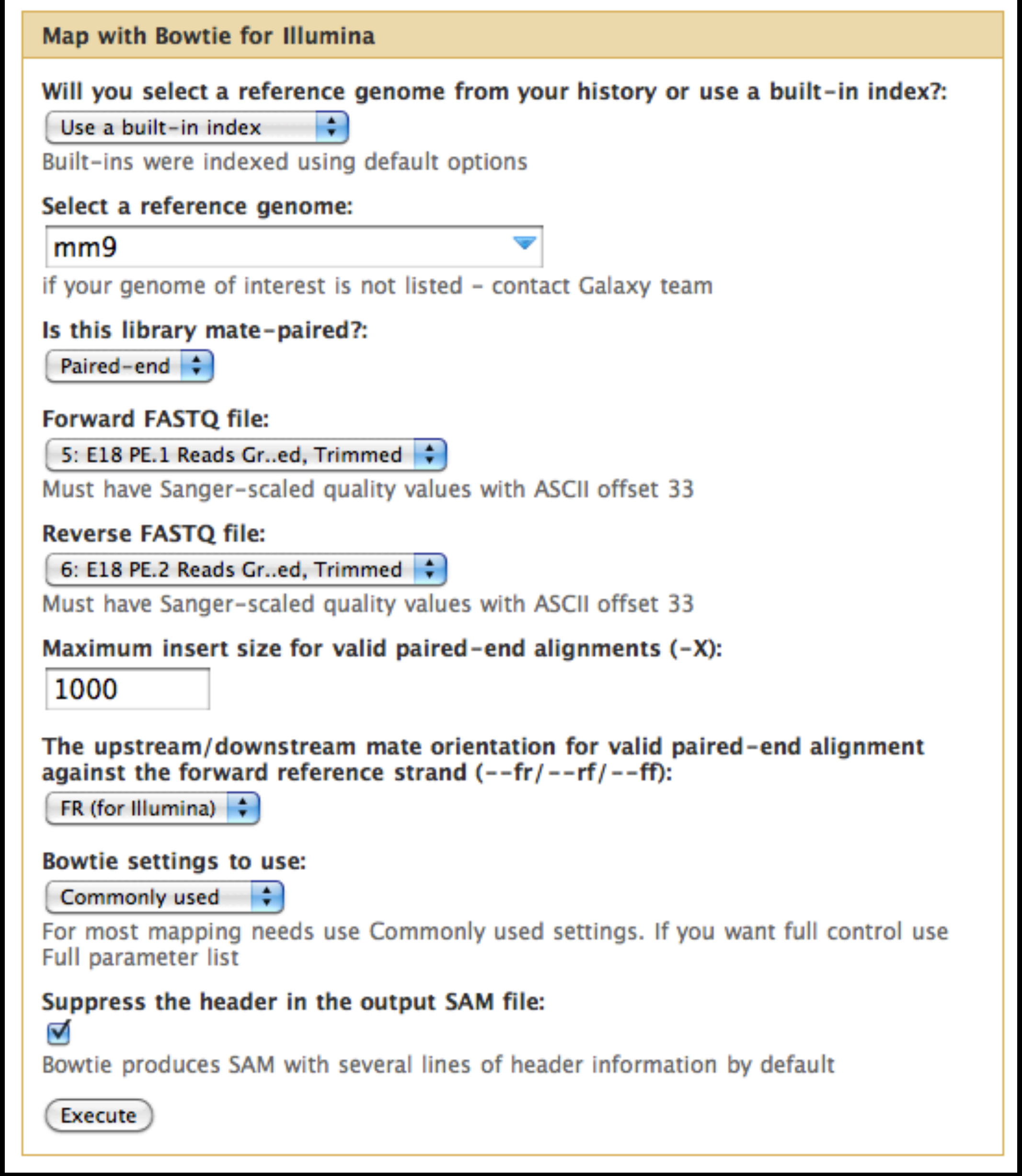


7: Map with Bowtie for Illumina on data 6 and data 5

9,073,928 lines, format: sam, database: mm9

info: Sequence file aligned.

1.QNAME	2.FLAG	3.I
HWI-EAS269:3:1:1449:913	99	chı
HWI-EAS269:3:1:1449:913	147	chı
HWI-EAS269:3:1:709:832	99	chı
HWI-EAS269:3:1:709:832	147	chı
HWI-EAS269:3:1:1422:1087	99	chı
HWI-EAS269:3:1:1422:1087	147	chı



Map with Bowtie for Illumina

Will you select a reference genome from your history or use a built-in index?: Use a built-in index

Built-ins were indexed using default options

Select a reference genome: mm9

if your genome of interest is not listed - contact Galaxy team

Is this library mate-paired?: Paired-end

Forward FASTQ file: 5: E18 PE.1 Reads Gr..ed, Trimmed

Must have Sanger-scaled quality values with ASCII offset 33

Reverse FASTQ file: 6: E18 PE.2 Reads Gr..ed, Trimmed

Must have Sanger-scaled quality values with ASCII offset 33

Maximum insert size for valid paired-end alignments (-X): 1000

The upstream/downstream mate orientation for valid paired-end alignment against the forward reference strand (--fr/--rf/--ff): FR (for Illumina)

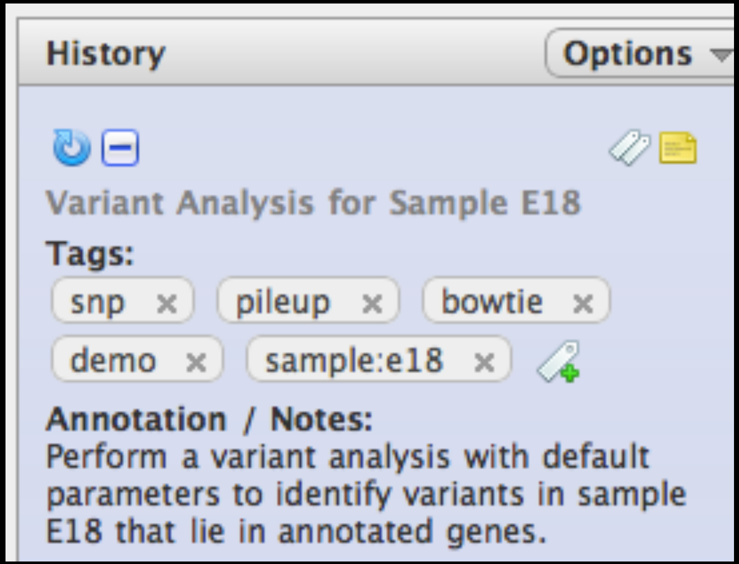
Bowtie settings to use: Commonly used

For most mapping needs use Commonly used settings. If you want full control use Full parameter list

Suppress the header in the output SAM file: ☑

Bowtie produces SAM with several lines of header information by default

As well as user-generated metadata and annotation...



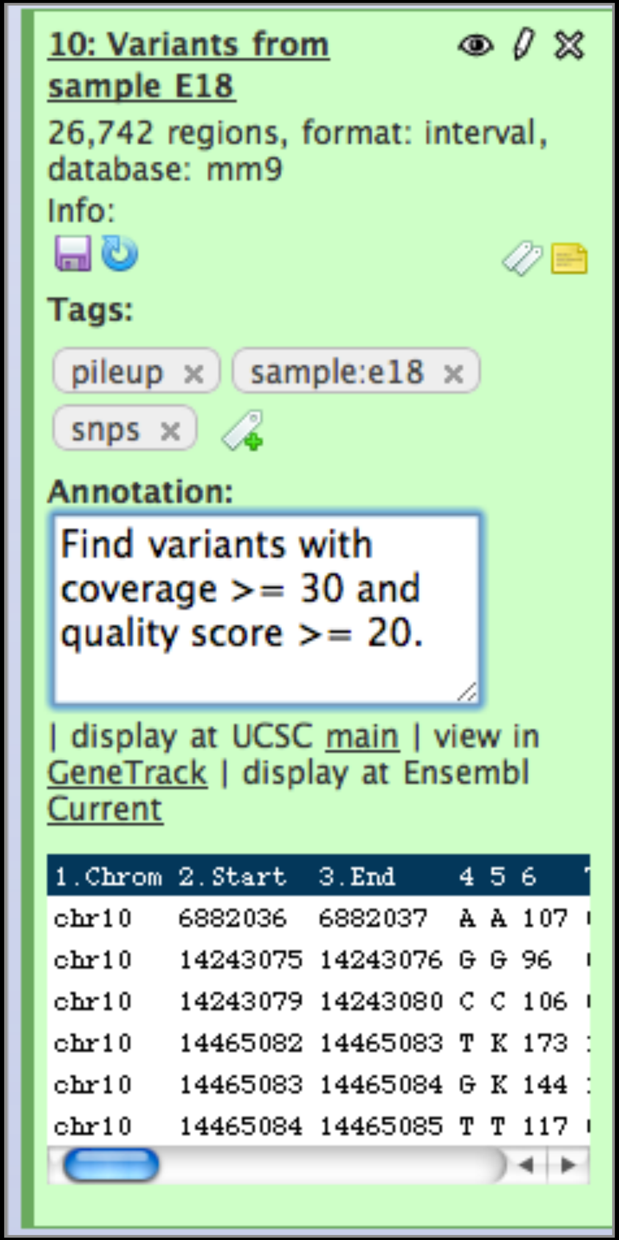
Variant Analysis for Sample E18

Tags:

snp × pileup × bowtie × demo × sample:e18 ×

Annotation / Notes:

Perform a variant analysis with default parameters to identify variants in sample E18 that lie in annotated genes.



10: Variants from sample E18

26,742 regions, format: interval, database: mm9

Info:

Tags:

pileup × sample:e18 × snps ×

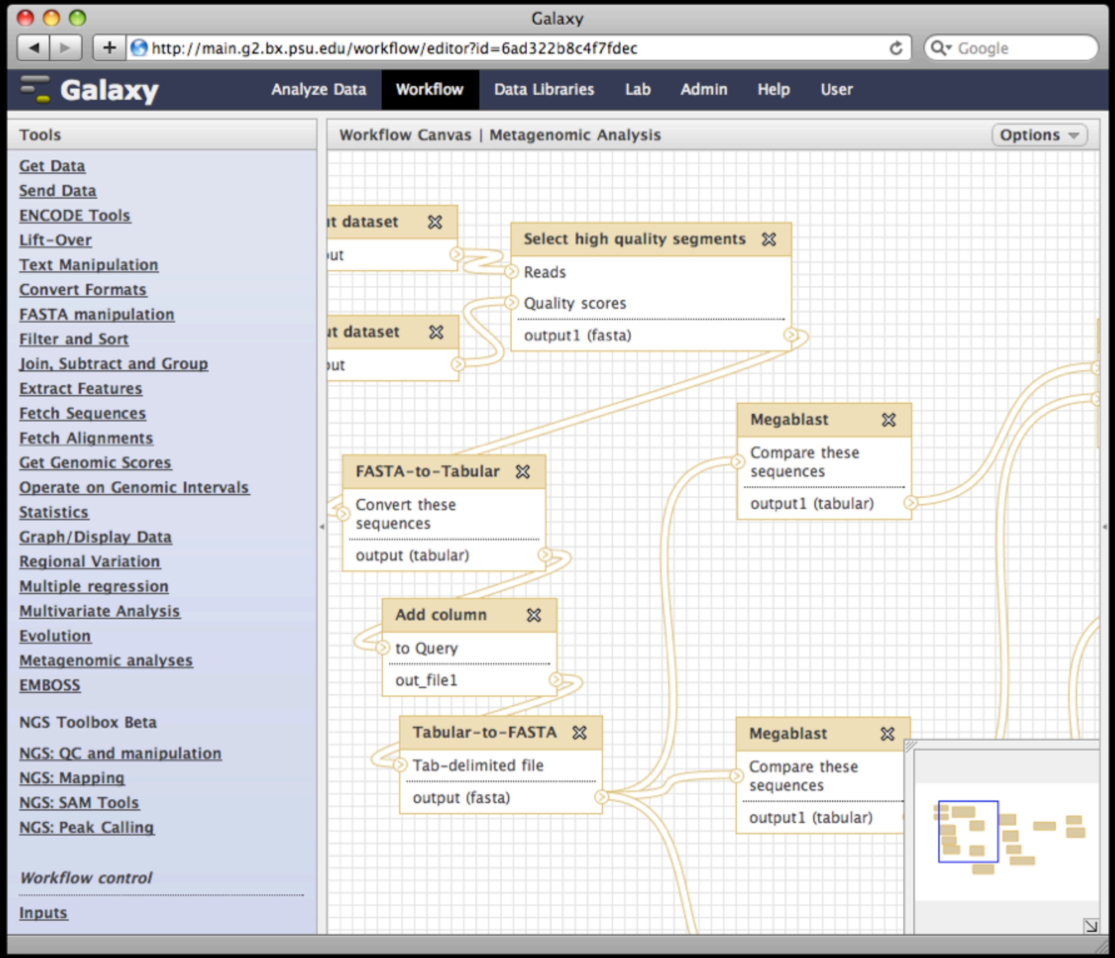
Annotation:

Find variants with coverage >= 30 and quality score >= 20.

| display at UCSC main | view in GeneTrack | display at Ensembl Current

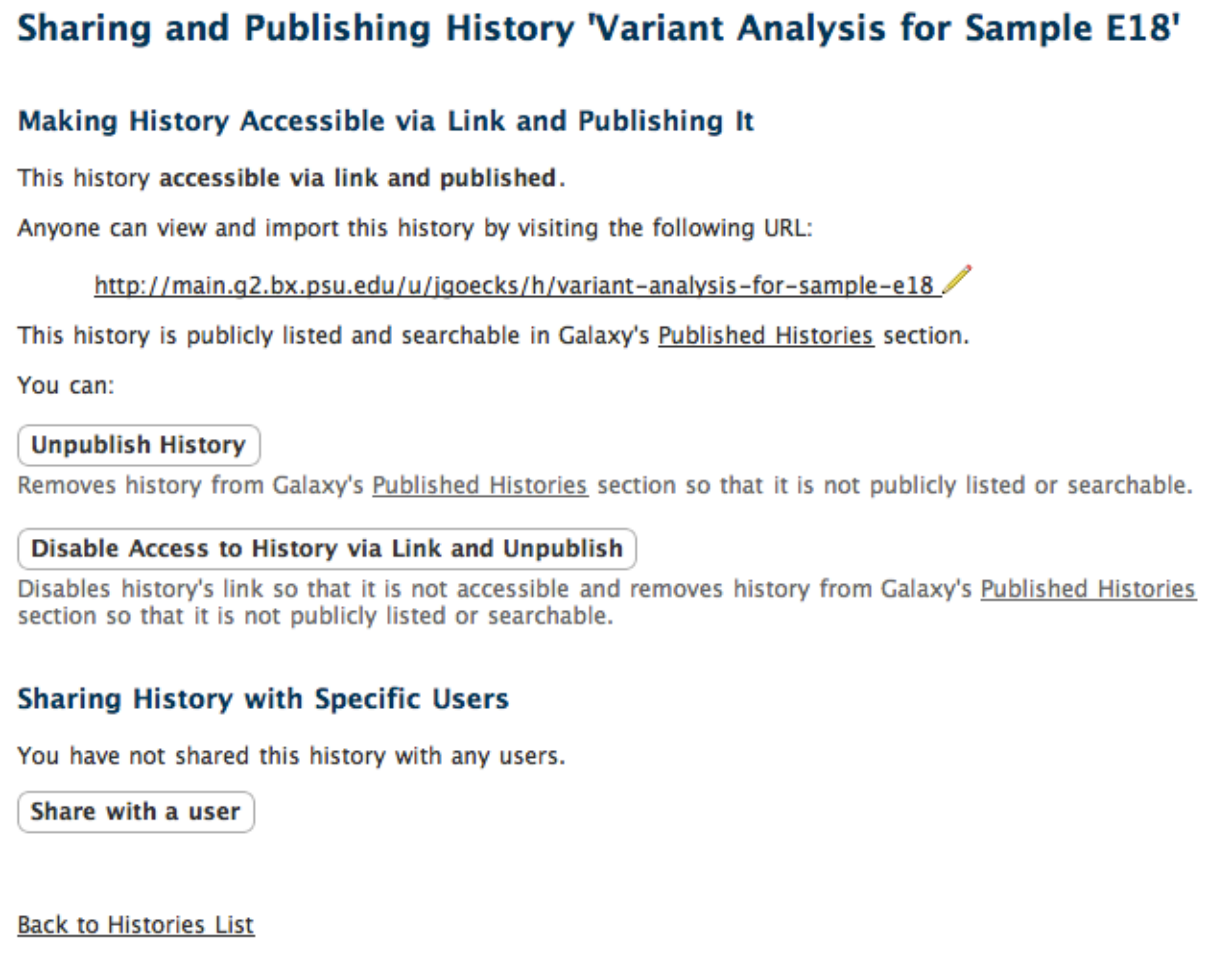
1.Chrom	2.Start	3.End	4	5	6
chr10	6882036	6882037	A	A	107
chr10	14243075	14243076	G	G	96
chr10	14243079	14243080	C	C	106
chr10	14465082	14465083	T	K	173
chr10	14465083	14465084	G	K	144
chr10	14465084	14465085	T	T	117

Galaxy workflow system



- Workflows can be constructed from scratch *or* extracted from existing analysis histories
- Facilitate reuse, as well as providing precise reproducibility of a complex analysis

Everything can be shared and published



Sharing and Publishing History 'Variant Analysis for Sample E18'

Making History Accessible via Link and Publishing It

This history **accessible via link and published**.

Anyone can view and import this history by visiting the following URL:

http://main.g2.bx.psu.edu/u/jgoecks/h/variant-analysis-for-sample-e18

This history is publicly listed and searchable in Galaxy's Published Histories section.

You can:

Removes history from Galaxy's Published Histories section so that it is not publicly listed or searchable.

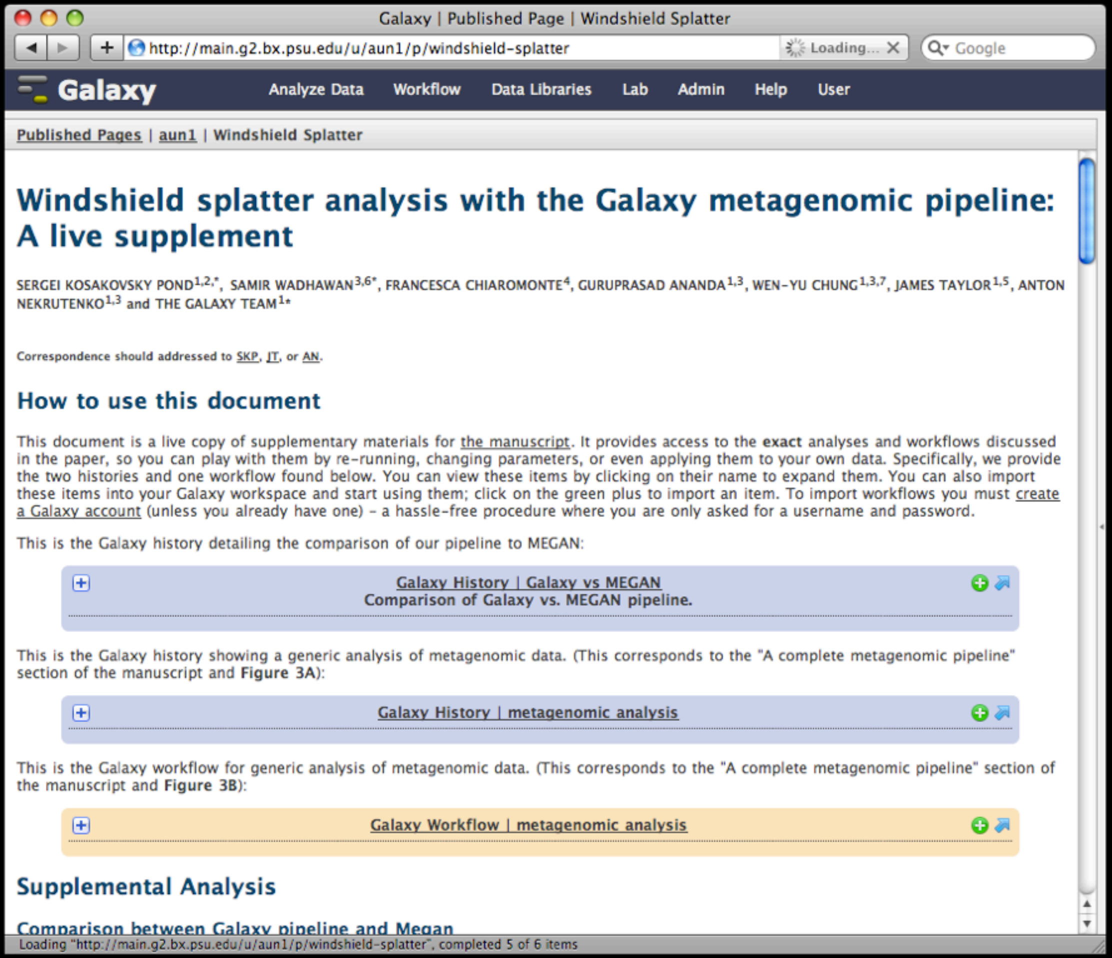
Disables history's link so that it is not accessible and removes history from Galaxy's Published Histories section so that it is not publicly listed or searchable.

Sharing History with Specific Users

You have not shared this history with any users.

Back to Histories List

Sharing and publishing


Galaxy | Published Page | Windshield Splatter
http://main.g2.bx.psu.edu/u/aun1/p/windshield-splatter
Loading...
Google
Galaxy
Analyze Data
Workflow
Data Libraries
Lab
Admin
Help
User
Published Pages | aun1 | Windshield Splatter

Windshield splatter analysis with the Galaxy metagenomic pipeline: A live supplement

SERGEI KOSAKOVSKY POND[1,2,*], SAMIR WADHAWAN[3,6*], FRANCESCA CHIAROMONTE[4], GURUPRASAD ANANDA[1,3], WEN-YU CHUNG[1,3,7], JAMES TAYLOR[1,5], ANTON NEKRUTENKO[1,3] and THE GALAXY TEAM[1*]

Correspondence should addressed to SKP, JT, or AN.

How to use this document

This document is a live copy of supplementary materials for the manuscript. It provides access to the **exact** analyses and workflows discussed in the paper, so you can play with them by re-running, changing parameters, or even applying them to your own data. Specifically, we provide the two histories and one workflow found below. You can view these items by clicking on their name to expand them. You can also import these items into your Galaxy workspace and start using them; click on the green plus to import an item. To import workflows you must create a Galaxy account (unless you already have one) - a hassle-free procedure where you are only asked for a username and password.

This is the Galaxy history detailing the comparison of our pipeline to MEGAN:

Galaxy History | Galaxy vs MEGAN
Comparison of Galaxy vs. MEGAN pipeline.

This is the Galaxy history showing a generic analysis of metagenomic data. (This corresponds to the "A complete metagenomic pipeline" section of the manuscript and **Figure 3A**):

Galaxy History | metagenomic analysis

This is the Galaxy workflow for generic analysis of metagenomic data. (This corresponds to the "A complete metagenomic pipeline" section of the manuscript and **Figure 3B**):

Galaxy Workflow | metagenomic analysis

Supplemental Analysis

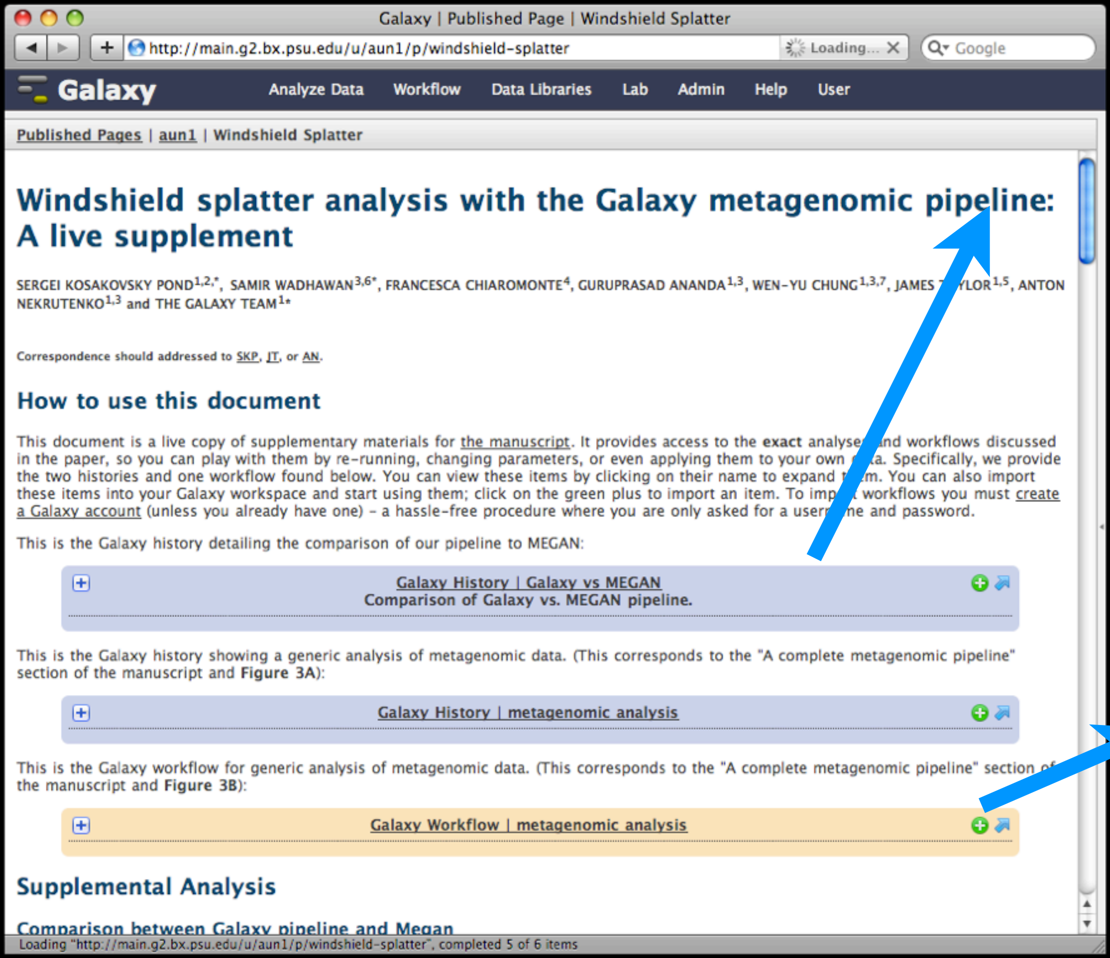
Comparison between Galaxy pipeline and Megan

Loading "http://main.g2.bx.psu.edu/u/aun1/p/windshield-splatter", completed 5 of 6 items

- All analysis components (datasets, histories, workflows) can be *shared* among Galaxy users and *published*

- Pages and annotation allow analaysis to be augmented with textual content and provided in the form of an integrated document

Sharing and publishing


Galaxy | Published Page | Windshield Splatter
http://main.g2.bx.psu.edu/u/aun1/p/windshield-splatter
Loading...
Google
Galaxy
Analyze Data
Workflow
Data Libraries
Lab
Admin
Help
User
Published Pages | aun1 | Windshield Splatter

Windshield splatter analysis with the Galaxy metagenomic pipeline: A live supplement

SERGEI KOSAKOVSKY POND[1,2,*], SAMIR WADHAWAN[3,6*], FRANCESCA CHIAROMONTE[4], GURUPRASAD ANANDA[1,3], WEN-YU CHUNG[1,3,7], JAMES TAYLOR[1,5], ANTON NEKRUTENKO[1,3] and THE GALAXY TEAM[1*]

Correspondence should addressed to SKP, JT, or AN.

How to use this document

This document is a live copy of supplementary materials for the manuscript. It provides access to the **exact** analyses and workflows discussed in the paper, so you can play with them by re-running, changing parameters, or even applying them to your own data. Specifically, we provide the two histories and one workflow found below. You can view these items by clicking on their name to expand them. You can also import these items into your Galaxy workspace and start using them; click on the green plus to import an item. To import workflows you must create a Galaxy account (unless you already have one) – a hassle-free procedure where you are only asked for a username and password.

This is the Galaxy history detailing the comparison of our pipeline to MEGAN:

Galaxy History | Galaxy vs MEGAN
Comparison of Galaxy vs. MEGAN pipeline.

This is the Galaxy history showing a generic analysis of metagenomic data. (This corresponds to the "A complete metagenomic pipeline" section of the manuscript and **Figure 3A**):

Galaxy History | metagenomic analysis

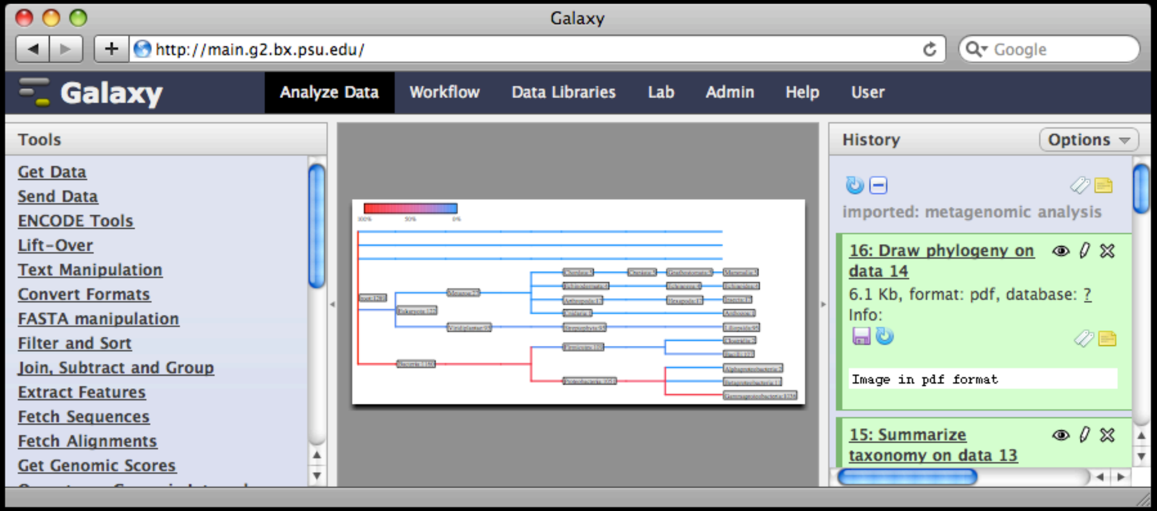
This is the Galaxy workflow for generic analysis of metagenomic data. (This corresponds to the "A complete metagenomic pipeline" section of the manuscript and **Figure 3B**):

Galaxy Workflow | metagenomic analysis

Supplemental Analysis

Comparison between Galaxy pipeline and Megan

Loading "http://main.g2.bx.psu.edu/u/aun1/p/windshield-splatter", completed 5 of 6 items

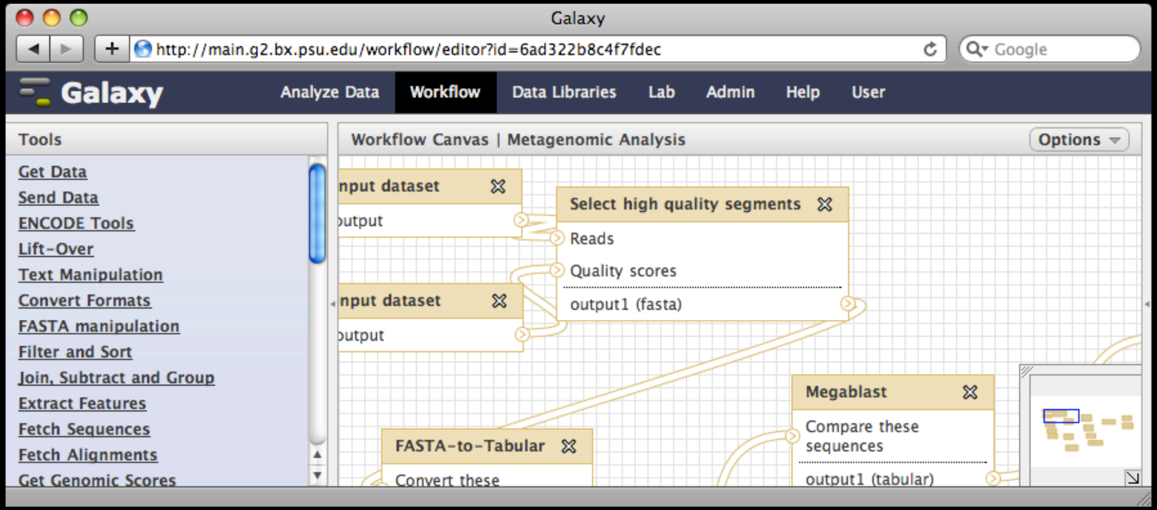


imported: metagenomic analysis

16: Draw phylogeny on data 14
6.1 Kb, format: pdf, database: ?
Info:

```
Image in pdf format
```

15: Summarize taxonomy on data 13



Making Galaxy your own

Building local Galaxy instances

- Galaxy is designed for local installation and customization
  - Just download and run, completely self-contained
  - Easily integrate new tools
  - Easy to deploy and manage on nearly any (unix) system
  - Run jobs on existing compute clusters

Scale up on your cluster

- Move intensive processing (tool execution) to other hosts
- Frees up the application server to serve requests and manage jobs
- Utilize existing resources
- Supports any scheduler that supports DRMAA (most of them)
- It's easy
- But, requires an **existing computational resource** on which to be deployed

CLUSTER RESOURCES INC.
TORQUE

GRID ENGINE

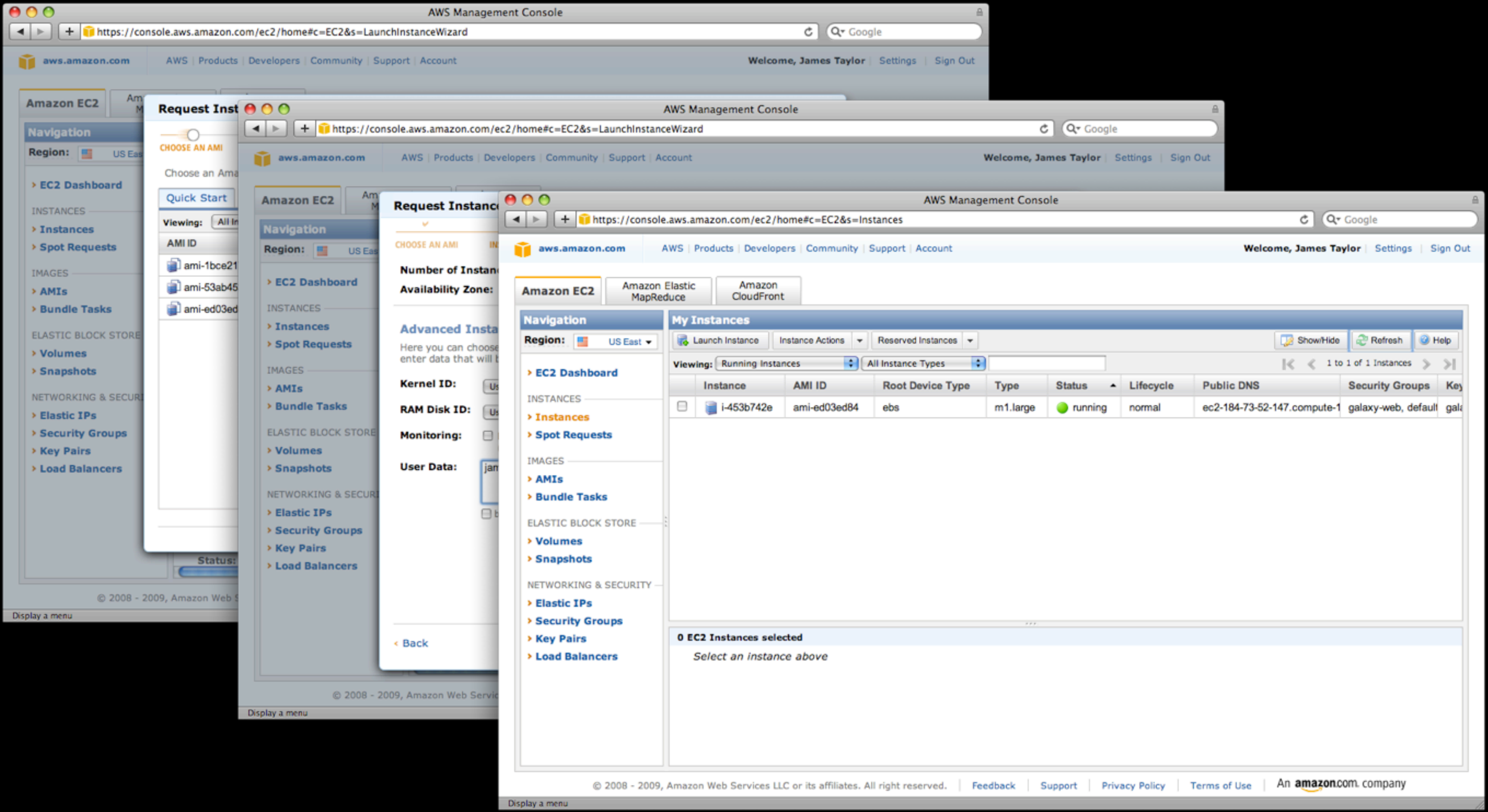
Platform Computing

DRMAA
Distributed Resource Management Application API — www.drmaa.org

Cloud computing / Infrastructure virtualization

- Computing using resources acquired on demand
- Virtual infrastructure allows for (potential) economies of scale, and (definite) improvements to management automation
- Cloud-style deployment provides a solution both for users without dedicated compute resources, and for simplifying deployment and management

Using Amazon EC2: Startup in 3 steps



AWS Management Console

https://console.aws.amazon.com/ec2/home#c=EC2&s=LaunchInstanceWizard

Request Inst

CHOOSE AN AMI

Choose an Ama

Quick Start

Viewing: All In

AMI ID
ami-1bce21
ami-53ab45
ami-ed03ed

AWS Management Console

https://console.aws.amazon.com/ec2/home#c=EC2&s=LaunchInstanceWizard

Request Instanc

CHOOSE AN AMI

Number of Instan

Availability Zone:

Advanced Insta

Here you can choose
enter data that will b

Kernel ID:

RAM Disk ID:

Monitoring: □

User Data: jam

‹ Back

AWS Management Console

https://console.aws.amazon.com/ec2/home#c=EC2&s=Instances

Welcome, James Taylor | Settings | Sign Out

My Instances

Region: US East

Viewing: Running Instances | All Instance Types

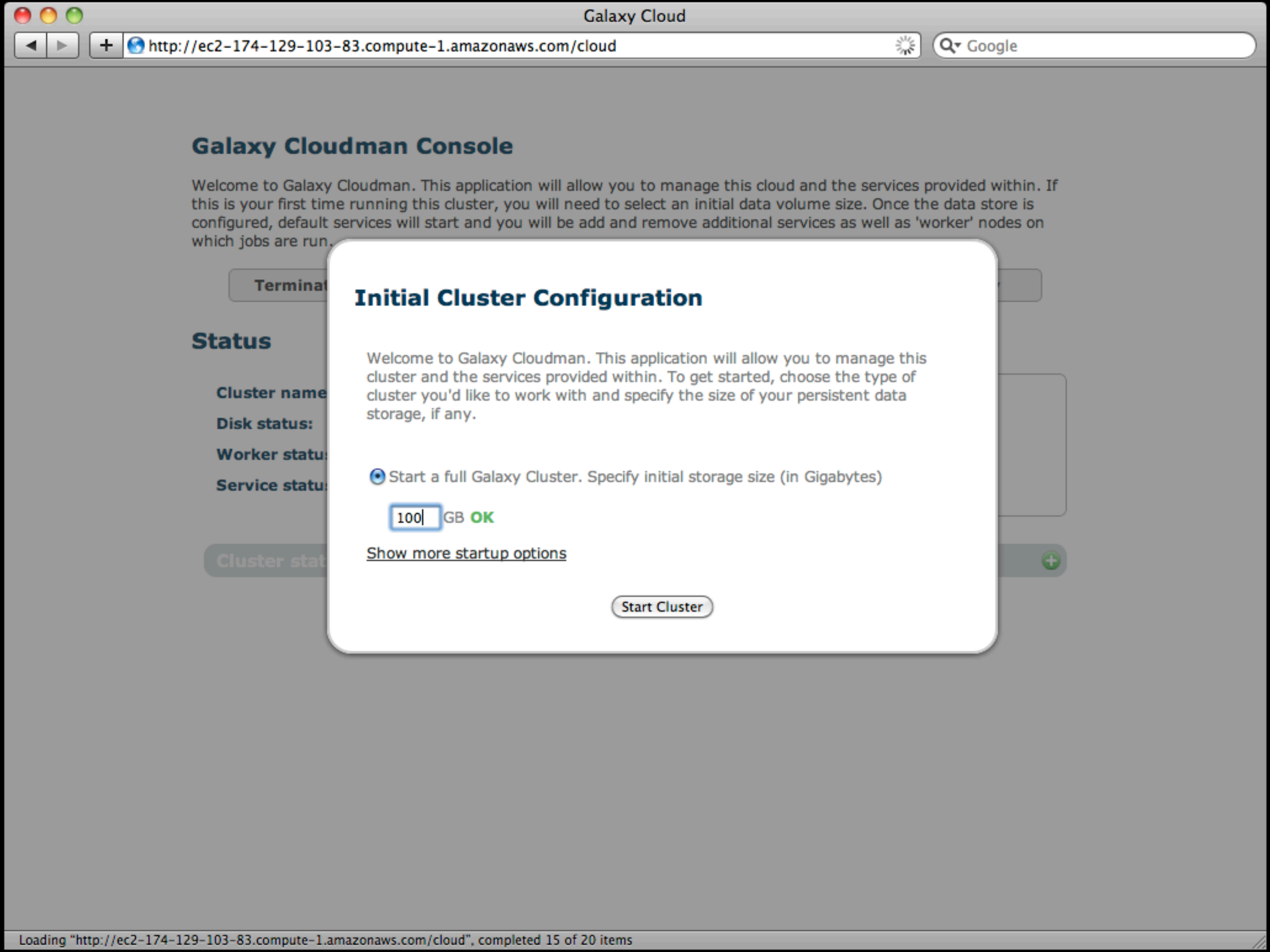
1 to 1 of 1 Instances

	Instance	AMI ID	Root Device Type	Type	Status	Lifecycle	Public DNS	Security Groups	Key
□	i-453b742e	ami-ed03ed84	ebs	m1.large	● running	normal	ec2-184-73-52-147.compute-1	galaxy-web, default	gala

0 EC2 Instances selected

Select an instance above

© 2008 - 2009, Amazon Web Services LLC or its affiliates. All right reserved.



Galaxy Cloudman Console

Welcome to Galaxy Cloudman. This application will allow you to manage this cloud and the services provided within. If this is your first time running this cluster, you will need to select an initial data volume size. Once the data store is configured, default services will start and you will be add and remove additional services as well as 'worker' nodes on which jobs are run.

Status

Initial Cluster Configuration

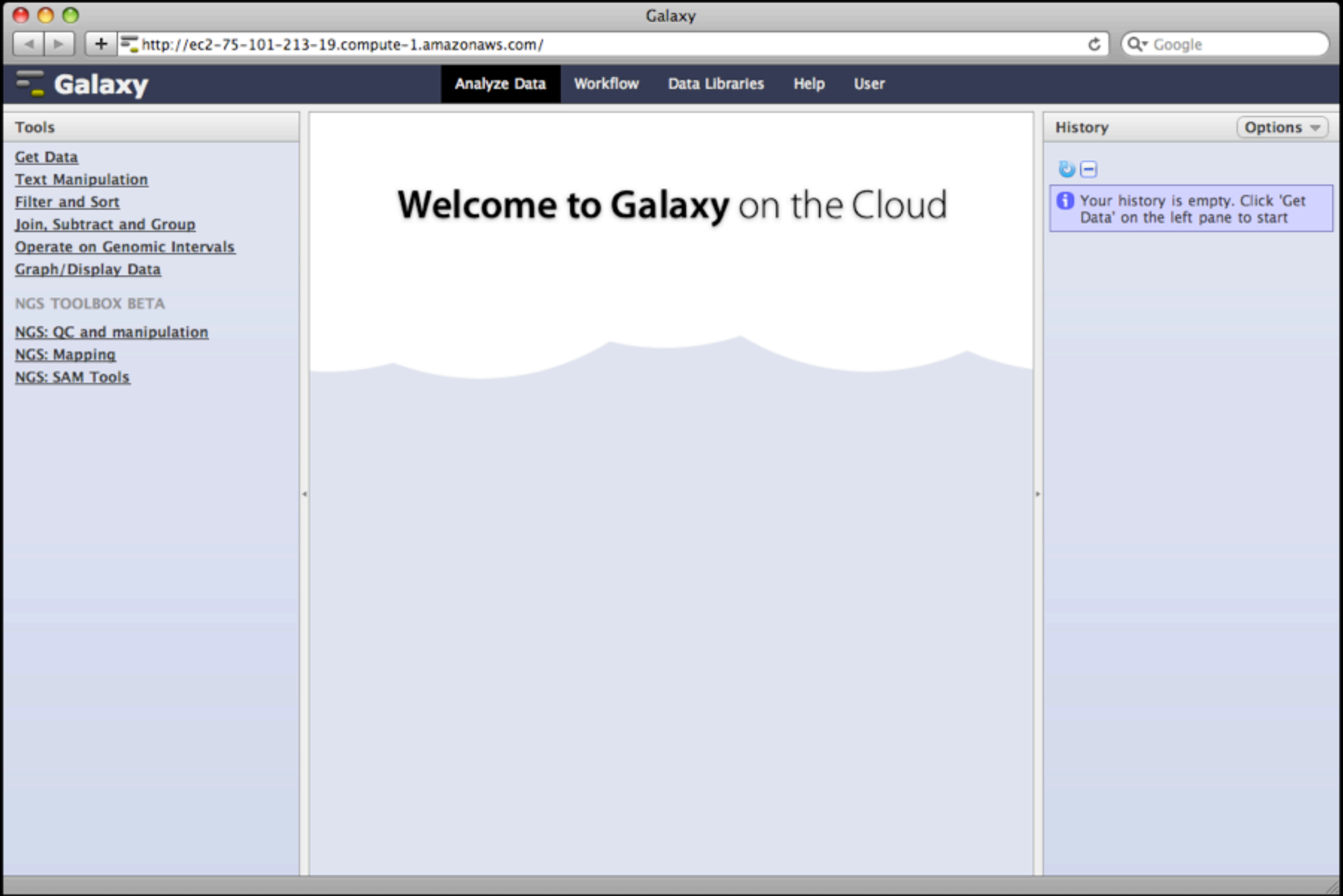
Welcome to Galaxy Cloudman. This application will allow you to manage this cluster and the services provided within. To get started, choose the type of cluster you'd like to work with and specify the size of your persistent data storage, if any.

● Start a full Galaxy Cluster. Specify initial storage size (in Gigabytes)

100 GB OK

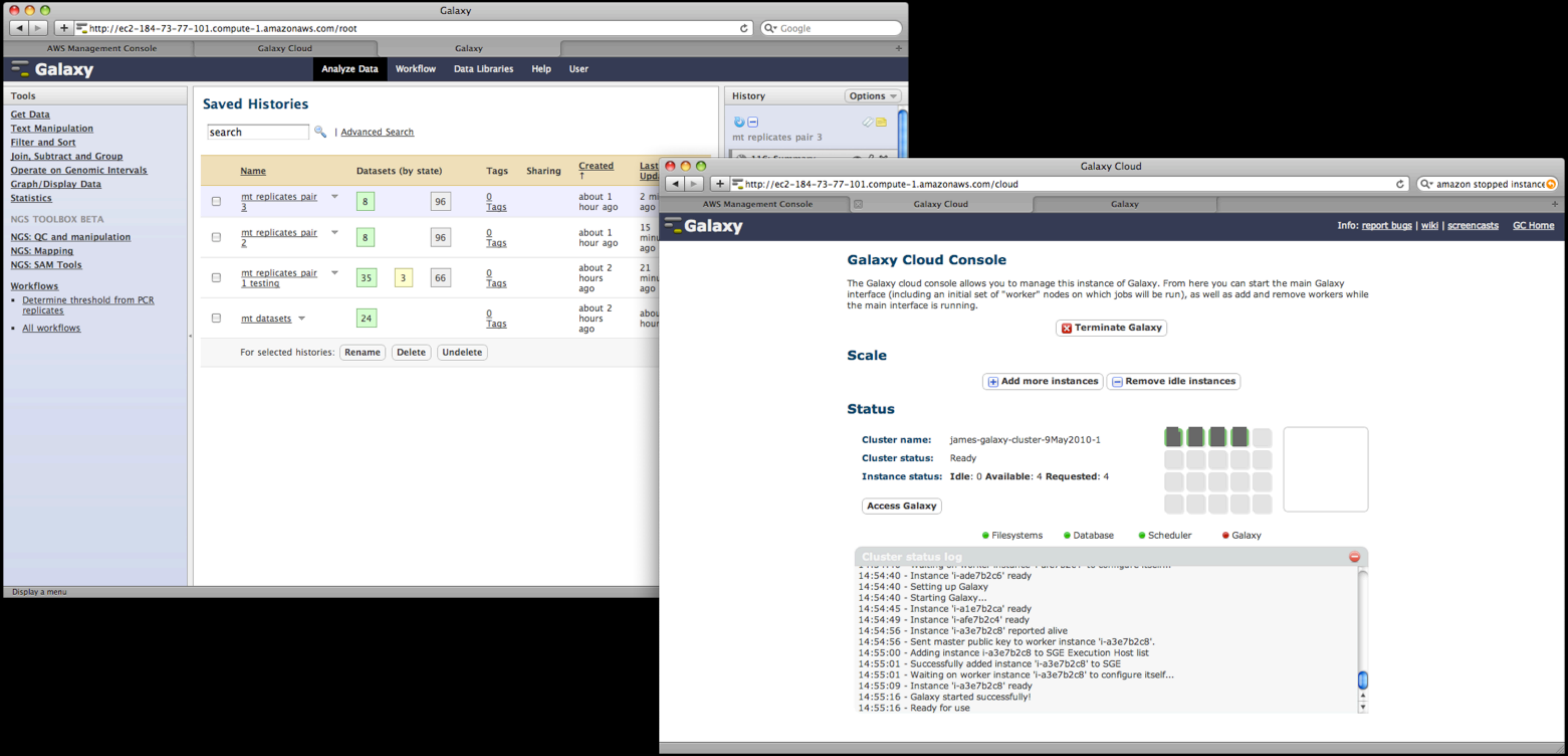
Show more startup options

Loading "http://ec2-174-129-103-83.compute-1.amazonaws.com/cloud", completed 15 of 20 items



Welcome to Galaxy on the Cloud

Your history is empty. Click 'Get Data' on the left pane to start



Saved Histories

search | Advanced Search

	Name	Datasets (by state)			Tags	Sharing	Created ↑	Last Upd
□	mt replicates pair 3	8		96	0 Tags		about 1 hour ago	2 mi ago
□	mt replicates pair 2	8		96	0 Tags		about 1 hour ago	15 minu ago
□	mt replicates pair 1 testing	35	3	66	0 Tags		about 2 hours ago	21 minu ago
□	mt datasets	24			0 Tags		about 2 hours ago	abou hour

For selected histories: Rename Delete Undelete

Galaxy Cloud
http://ec2-184-73-77-101.compute-1.amazonaws.com/cloud
amazon stopped instance
AWS Management Console
Galaxy Cloud
Galaxy
Galaxy
Info: report bugs | wiki | screencasts
GC Home

Galaxy Cloud Console

The Galaxy cloud console allows you to manage this instance of Galaxy. From here you can start the main Galaxy interface (including an initial set of "worker" nodes on which jobs will be run), as well as add and remove workers while the main interface is running.

Scale

Status

Cluster name: james-galaxy-cluster-9May2010-1

Cluster status: Ready

Instance status: **Idle**: 0 **Available**: 4 **Requested**: 4

Cluster status log

```
14:54:40 - Instance 'i-ade7b2c6' ready
14:54:40 - Setting up Galaxy
14:54:40 - Starting Galaxy...
14:54:45 - Instance 'i-a1e7b2ca' ready
14:54:49 - Instance 'i-afe7b2c4' ready
14:54:56 - Instance 'i-a3e7b2c8' reported alive
14:54:56 - Sent master public key to worker instance 'i-a3e7b2c8'.
14:55:00 - Adding instance i-a3e7b2c8 to SGE Execution Host list
14:55:01 - Successfully added instance 'i-a3e7b2c8' to SGE
14:55:01 - Waiting on worker instance 'i-a3e7b2c8' to configure itself...
14:55:09 - Instance 'i-a3e7b2c8' ready
14:55:16 - Galaxy started successfully!
14:55:16 - Ready for use
```

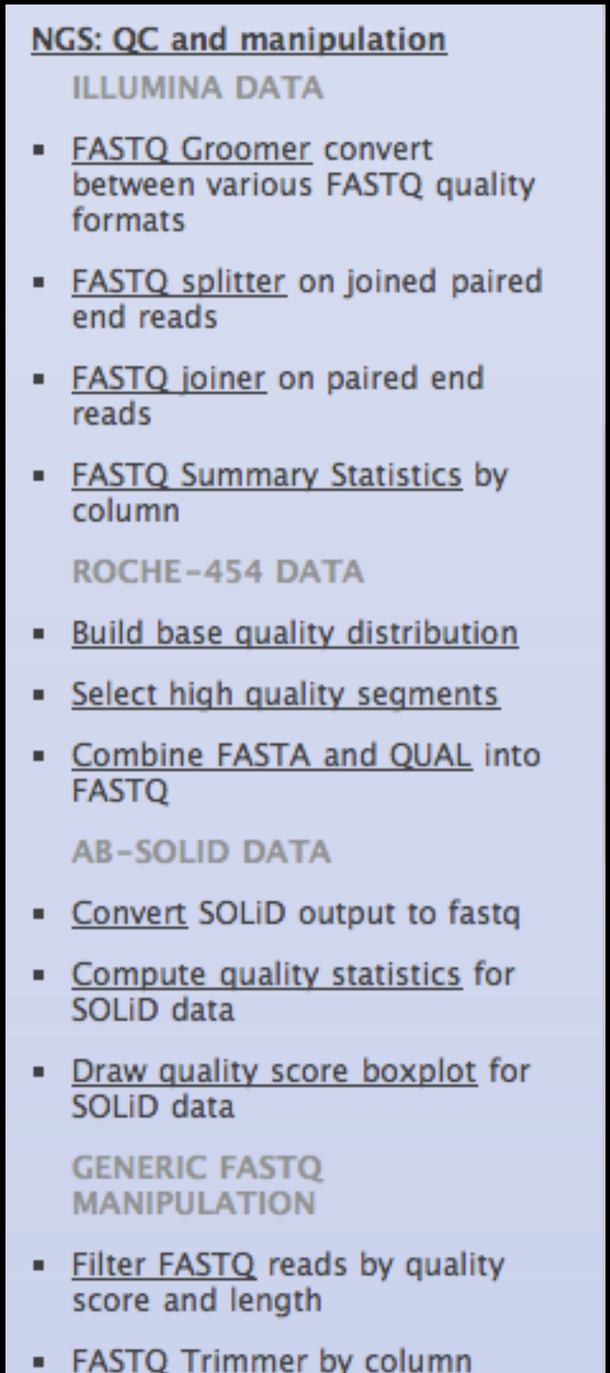
Can use like any other Galaxy instance, with additional compute nodes acquired and released (*automatically*) in response to usage

Analyzing high throughput sequence data with Galaxy

- The Galaxy framework is generic, supporting a new type of analysis is as simple as integrating tools
- Galaxy is well suited to large-scale analysis
  - Allows tools to work with data in native, efficient formats
  - Integrates easily with cluster computing resources

http://usegalaxy.org/heteroplasmy

(some) Galaxy tools for sequence data analysis



NGS: QC and manipulation

ILLUMINA DATA

- FASTQ Groomer convert between various FASTQ quality formats
- FASTQ splitter on joined paired end reads
- FASTQ joiner on paired end reads
- FASTQ Summary Statistics by column

ROCHE-454 DATA

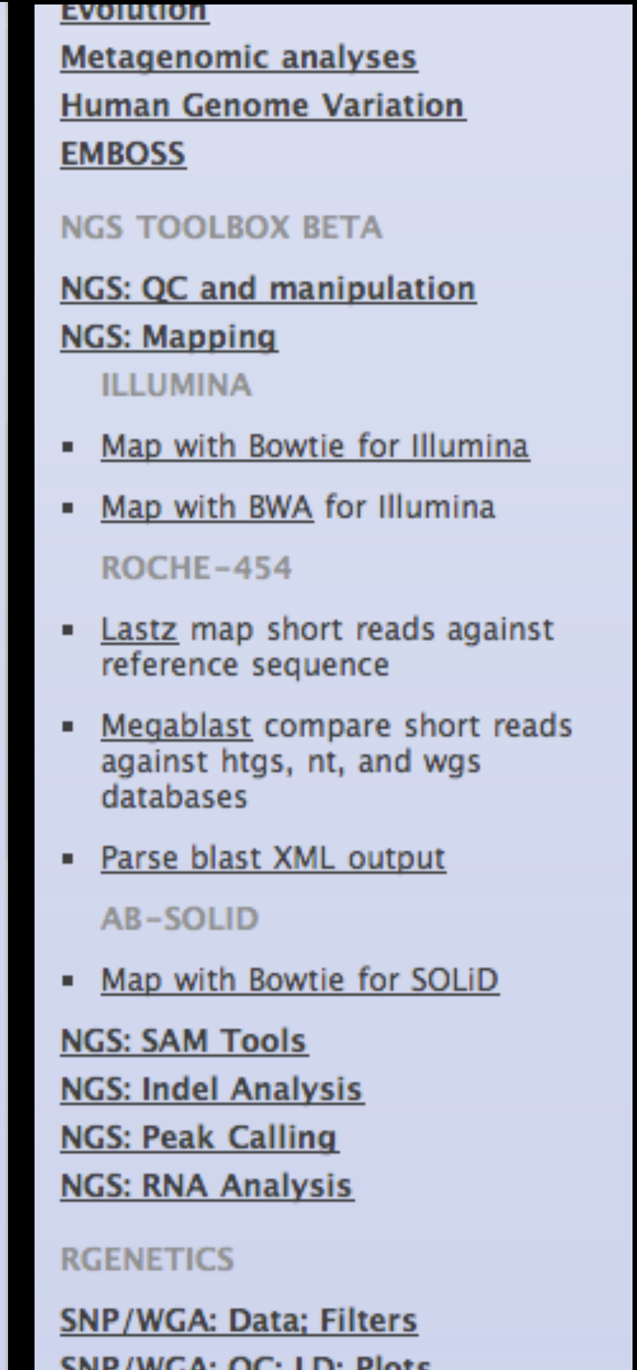
- Build base quality distribution
- Select high quality segments
- Combine FASTA and QUAL into FASTQ

AB-SOLID DATA

- Convert SOLiD output to fastq
- Compute quality statistics for SOLiD data
- Draw quality score boxplot for SOLiD data

GENERIC FASTQ MANIPULATION

- Filter FASTQ reads by quality score and length
- FASTQ Trimmer by column



Evolution

Metagenomic analyses

Human Genome Variation

EMBOSS

NGS TOOLBOX BETA

NGS: QC and manipulation

NGS: Mapping

ILLUMINA

- Map with Bowtie for Illumina
- Map with BWA for Illumina

ROCHE-454

- Lastz map short reads against reference sequence
- Megablast compare short reads against htgs, nt, and wgs databases
- Parse blast XML output

AB-SOLID

- Map with Bowtie for SOLiD

NGS: SAM Tools

NGS: Indel Analysis

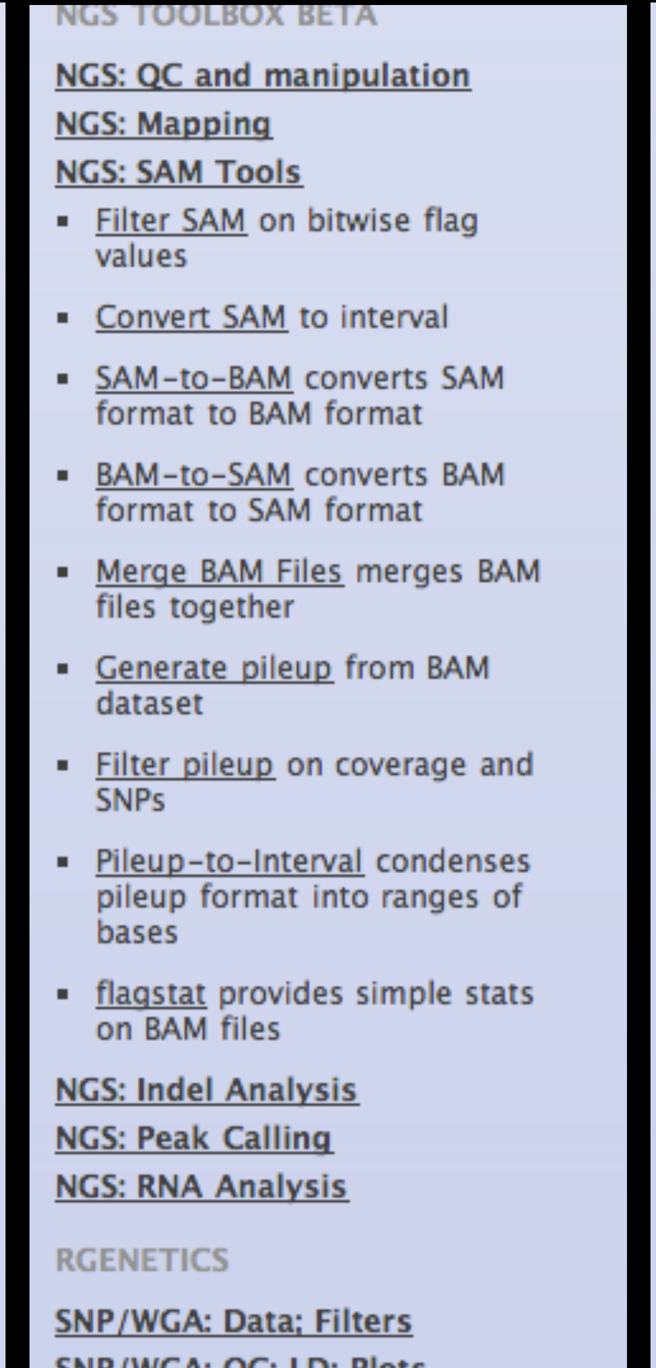
NGS: Peak Calling

NGS: RNA Analysis

RGENETICS

SNP/WGA: Data; Filters

SNP/WGA: QC; LD; Plots



NGS TOOLBOX BETA

NGS: QC and manipulation

NGS: Mapping

NGS: SAM Tools

- Filter SAM on bitwise flag values
- Convert SAM to interval
- SAM-to-BAM converts SAM format to BAM format
- BAM-to-SAM converts BAM format to SAM format
- Merge BAM Files merges BAM files together
- Generate pileup from BAM dataset
- Filter pileup on coverage and SNPs
- Pileup-to-Interval condenses pileup format into ranges of bases
- flagstat provides simple stats on BAM files

NGS: Indel Analysis

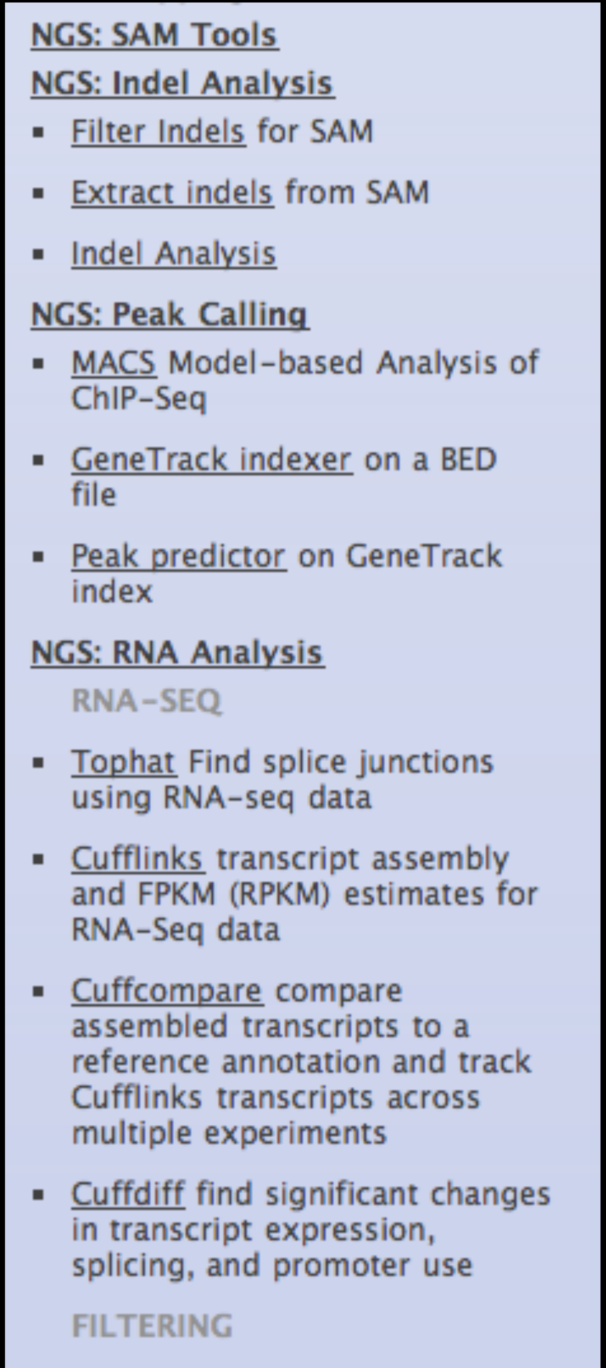
NGS: Peak Calling

NGS: RNA Analysis

RGENETICS

SNP/WGA: Data; Filters

SNP/WGA: QC; LD; Plots



NGS: SAM Tools

NGS: Indel Analysis

- Filter Indels for SAM
- Extract indels from SAM
- Indel Analysis

NGS: Peak Calling

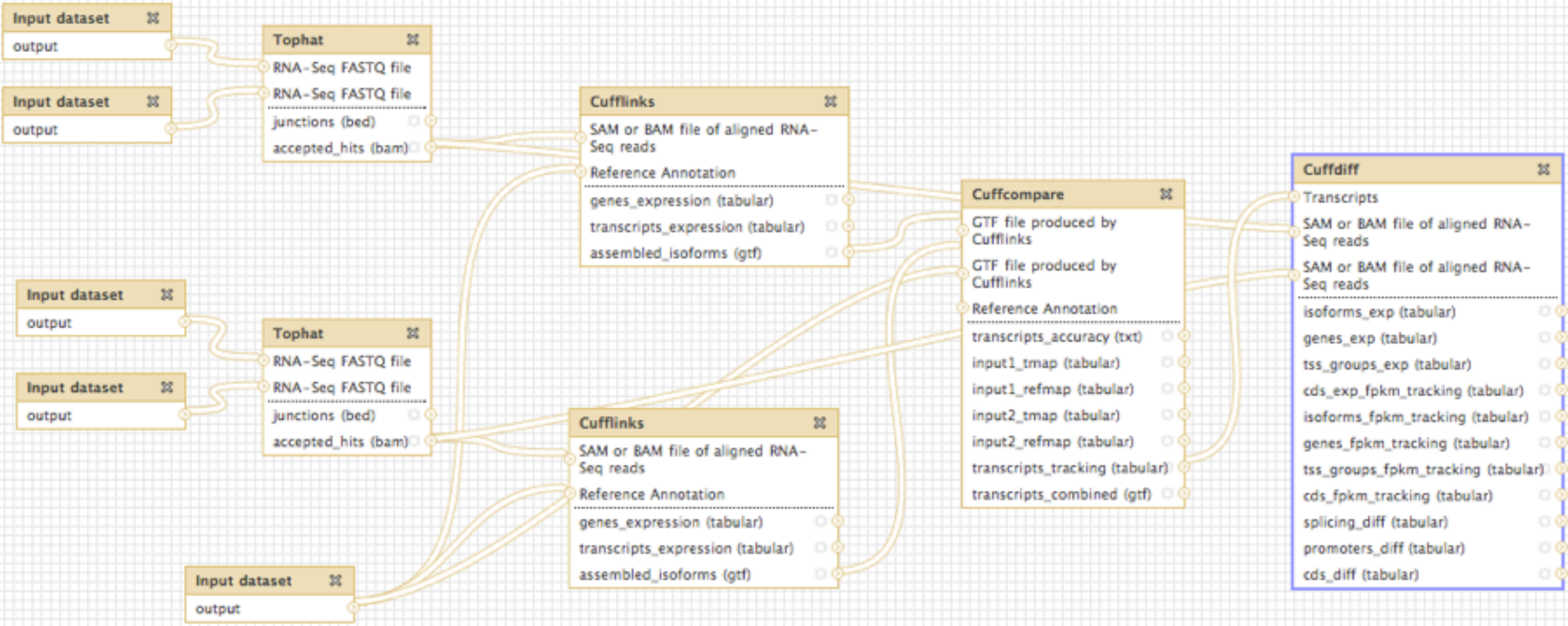
- MACS Model-based Analysis of ChIP-Seq
- GeneTrack indexer on a BED file
- Peak predictor on GeneTrack index

NGS: RNA Analysis

RNA-SEQ

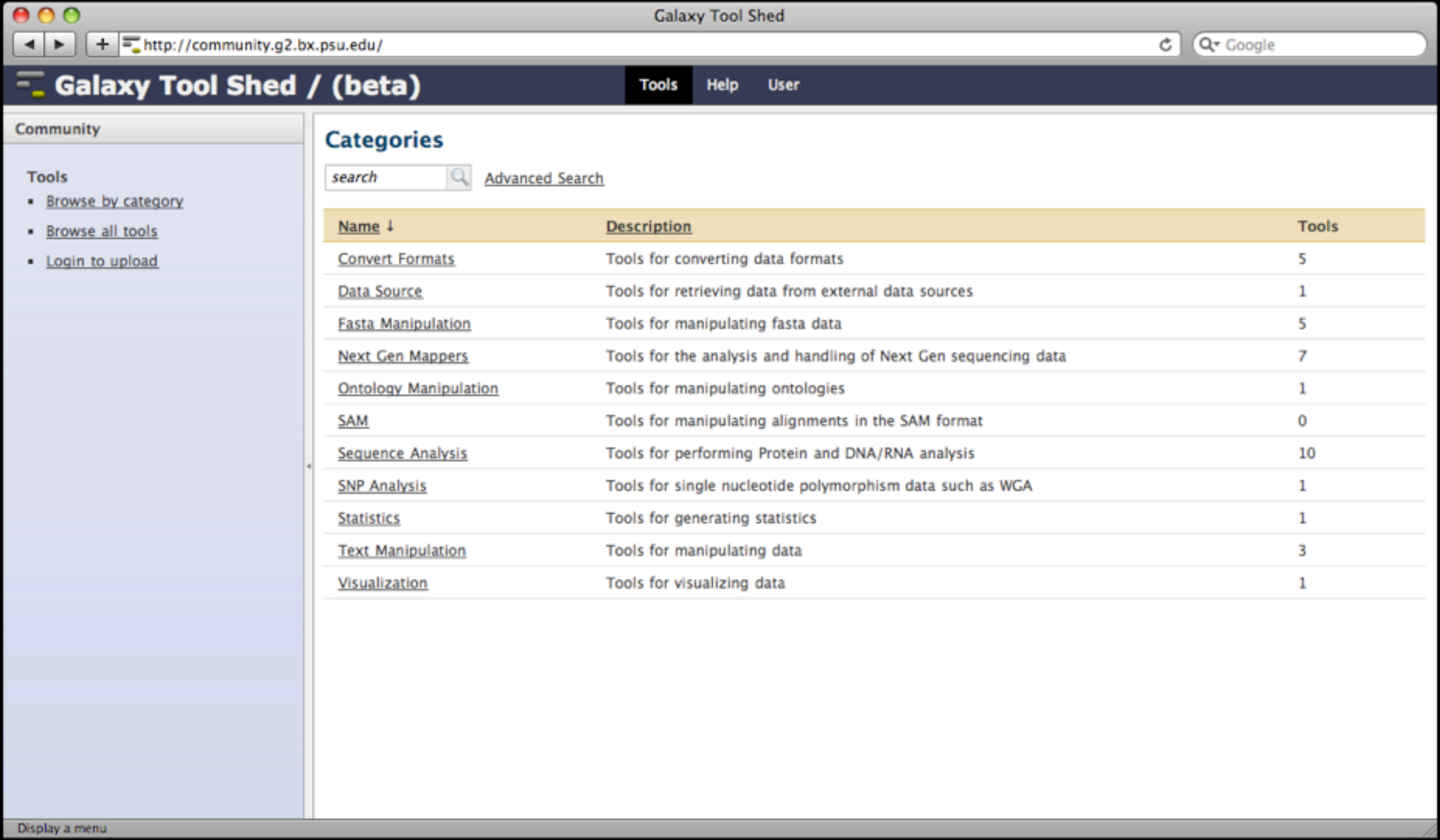
- Tophat Find splice junctions using RNA-seq data
- Cufflinks transcript assembly and FPKM (RPKM) estimates for RNA-Seq data
- Cuffcompare compare assembled transcripts to a reference annotation and track Cufflinks transcripts across multiple experiments
- Cuffdiff find significant changes in transcript expression, splicing, and promoter use

FILTERING



Example: Workflow for differential expression analysis of RNA-seq using Tophat/Cufflinks tools

Community of tool developers

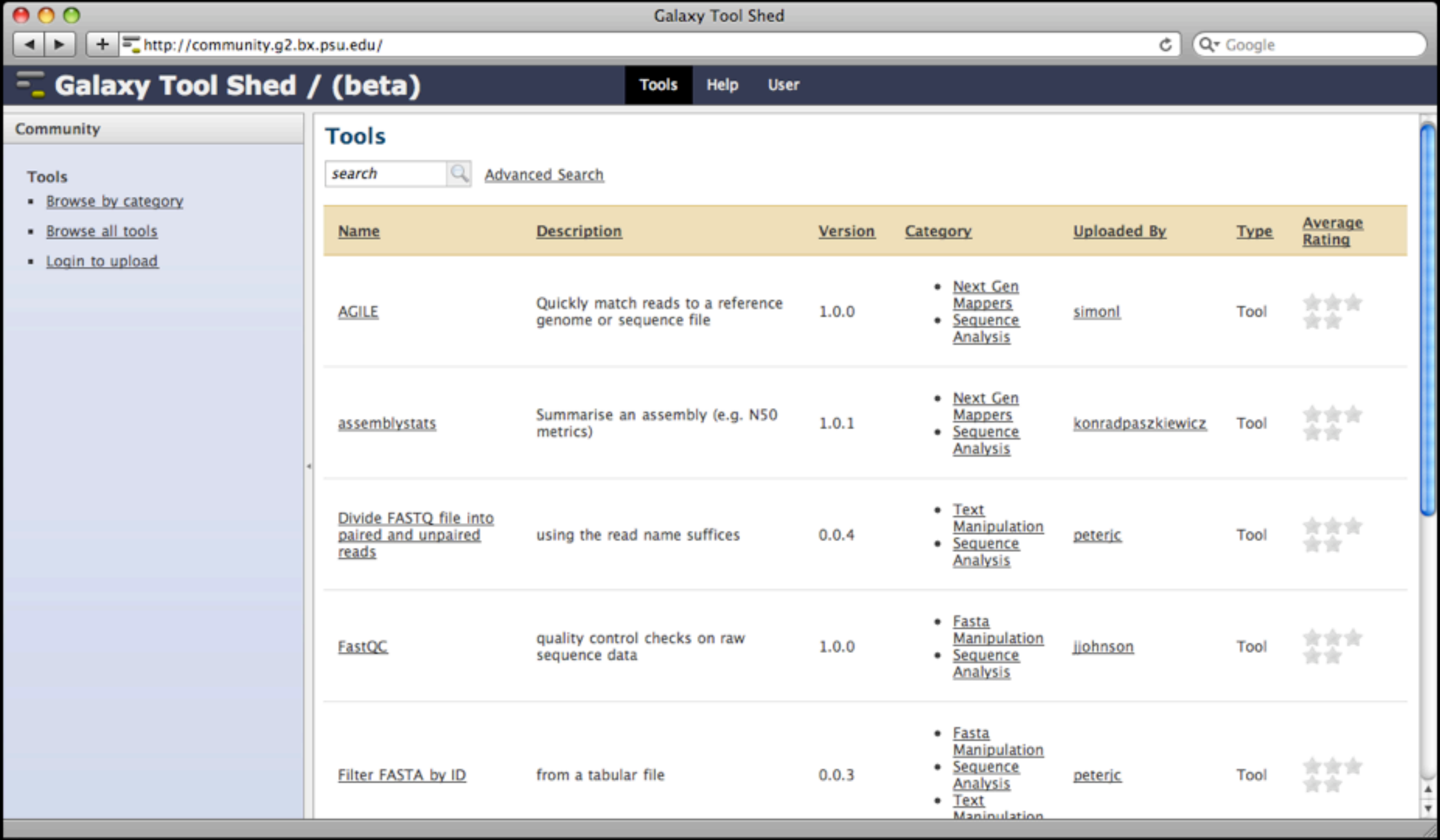


Tools

- Browse by category
- Browse all tools
- Login to upload

Categories

Name ↓	Description	Tools
Convert Formats	Tools for converting data formats	5
Data Source	Tools for retrieving data from external data sources	1
Fasta Manipulation	Tools for manipulating fasta data	5
Next Gen Mappers	Tools for the analysis and handling of Next Gen sequencing data	7
Ontology Manipulation	Tools for manipulating ontologies	1
SAM	Tools for manipulating alignments in the SAM format	0
Sequence Analysis	Tools for performing Protein and DNA/RNA analysis	10
SNP Analysis	Tools for single nucleotide polymorphism data such as WGA	1
Statistics	Tools for generating statistics	1
Text Manipulation	Tools for manipulating data	3
Visualization	Tools for visualizing data	1

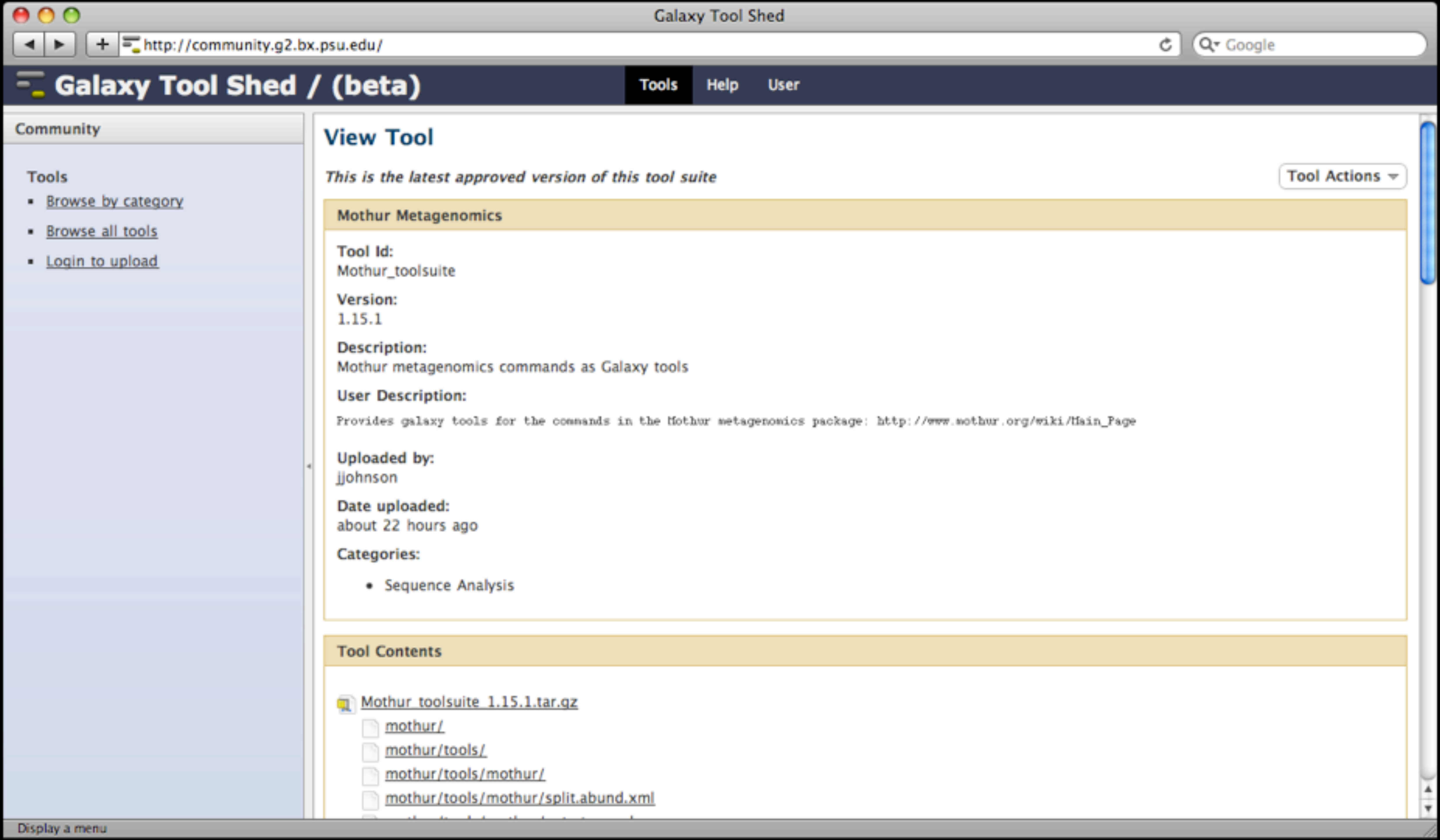


Tools

- Browse by category
- Browse all tools
- Login to upload

Tools

Name	Description	Version	Category	Uploaded By	Type	Average Rating
AGILE	Quickly match reads to a reference genome or sequence file	1.0.0	• Next Gen Mappers • Sequence Analysis	simonl	Tool	☆☆☆☆☆
assemblystats	Summarise an assembly (e.g. N50 metrics)	1.0.1	• Next Gen Mappers • Sequence Analysis	konradpaszkiewicz	Tool	☆☆☆☆☆
Divide FASTQ file into paired and unpaired reads	using the read name suffices	0.0.4	• Text Manipulation • Sequence Analysis	peterjc	Tool	☆☆☆☆☆
FastQC	quality control checks on raw sequence data	1.0.0	• Fasta Manipulation • Sequence Analysis	jjohnson	Tool	☆☆☆☆☆
Filter FASTA by ID	from a tabular file	0.0.3	• Fasta Manipulation • Sequence Analysis • Text Manipulation	peterjc	Tool	☆☆☆☆☆



View Tool

This is the latest approved version of this tool suite

Mothur Metagenomics

Tool Id:
Mothur_toolsuite

Version:
1.15.1

Description:
Mothur metagenomics commands as Galaxy tools

User Description:

```
Provides galaxy tools for the commands in the Mothur metagenomics package: http://www.mothur.org/wiki/Main_Page
```

Uploaded by:
jjohnson

Date uploaded:
about 22 hours ago

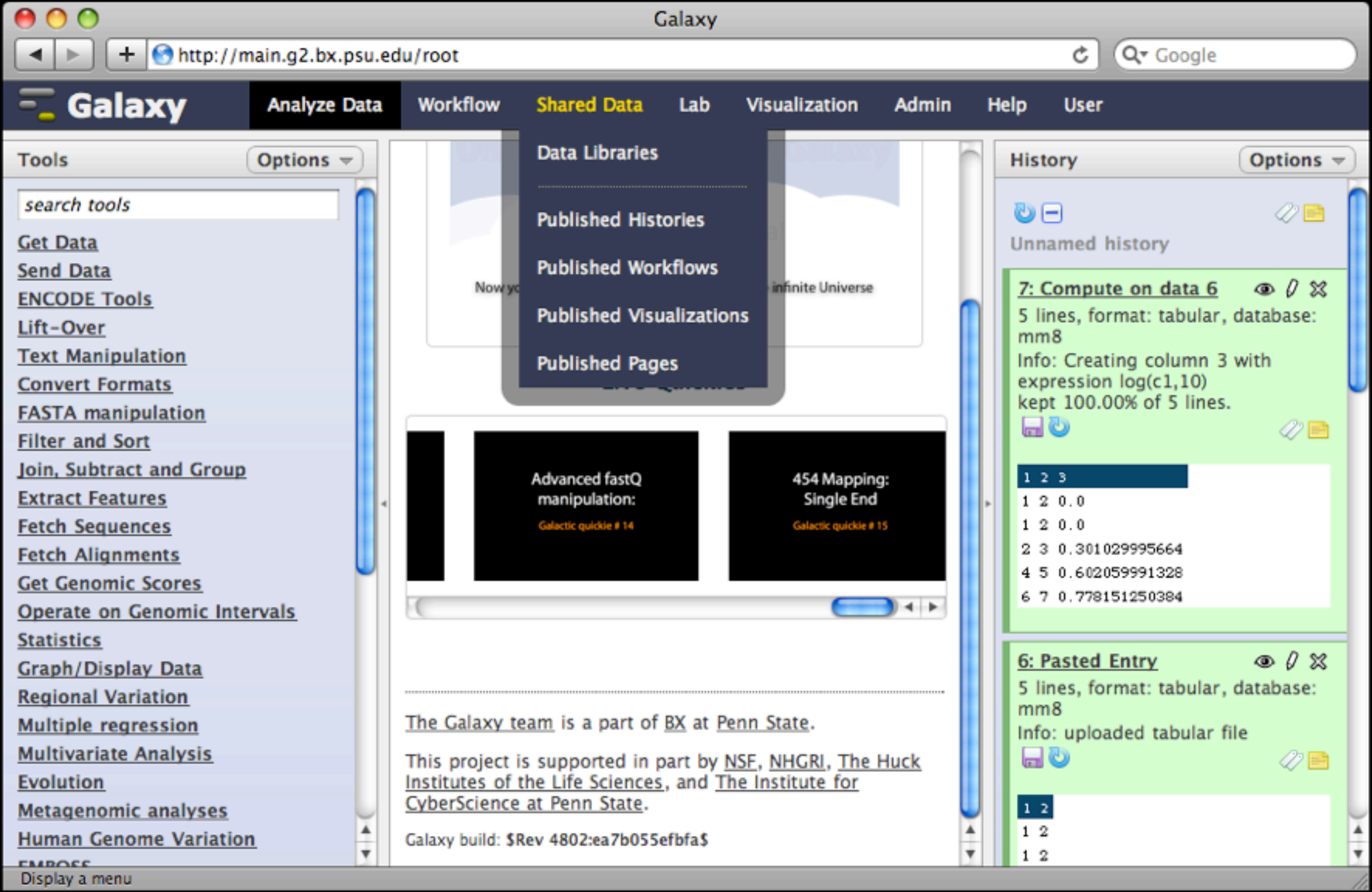
Categories:

- Sequence Analysis

Tool Contents

Mothur_toolsuite_1.15.1.tar.gz
mothur/
mothur/tools/
mothur/tools/mothur/
mothur/tools/mothur/split.abund.xml

Data management



7: Compute on data 6

5 lines, format: tabular, database: mm8

Info: Creating column 3 with expression log(c1,10)
kept 100.00% of 5 lines.

```
1 2 3
1 2 0.0
1 2 0.0
2 3 0.301029995664
4 5 0.602059991328
6 7 0.778151250384
```

6: Pasted Entry

5 lines, format: tabular, database: mm8

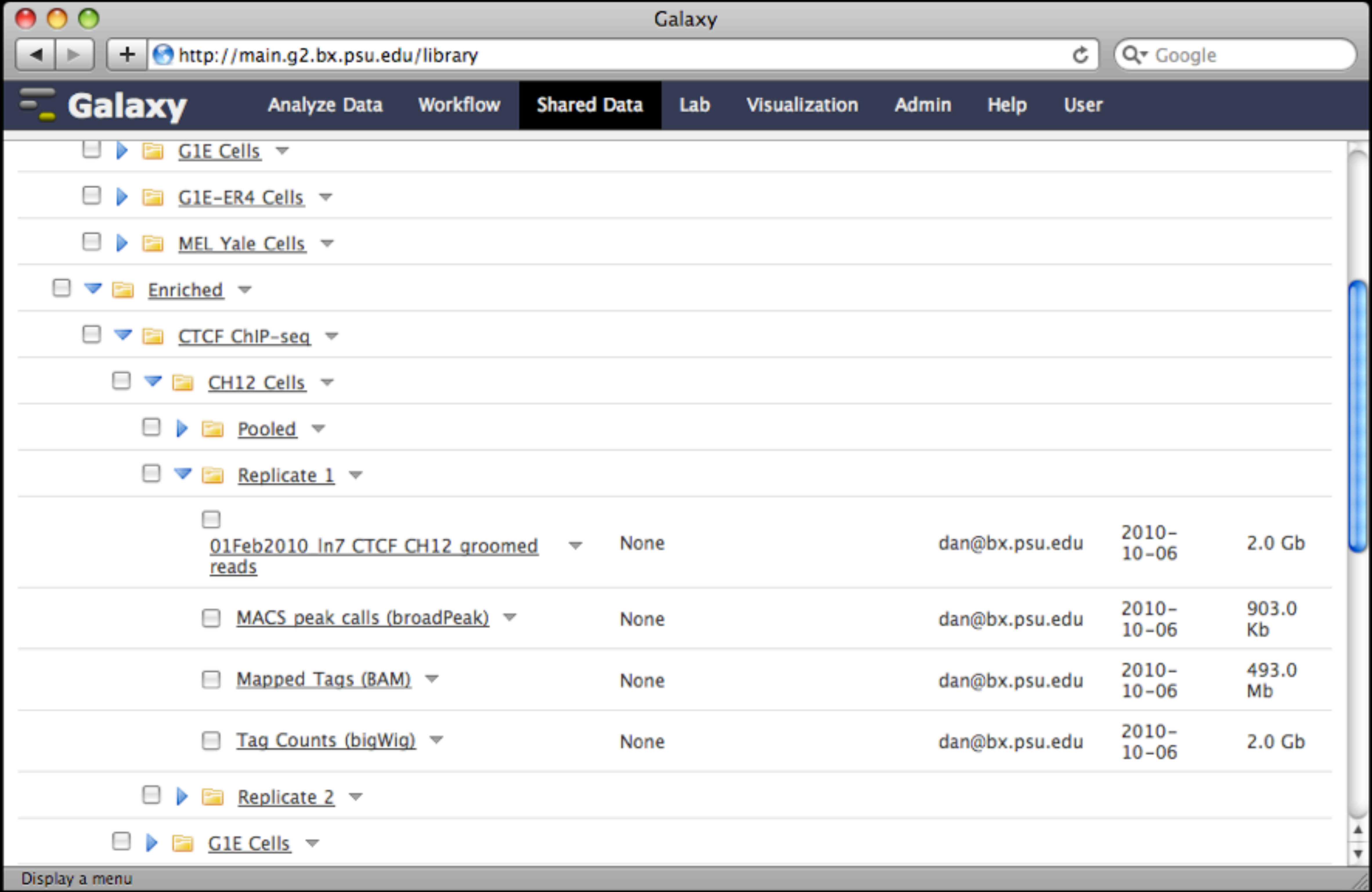
Info: uploaded tabular file

```
1 2
1 2
1 2
```

The Galaxy team is a part of BX at Penn State.

This project is supported in part by NSF, NHGRI, The Huck Institutes of the Life Sciences, and The Institute for CyberScience at Penn State.

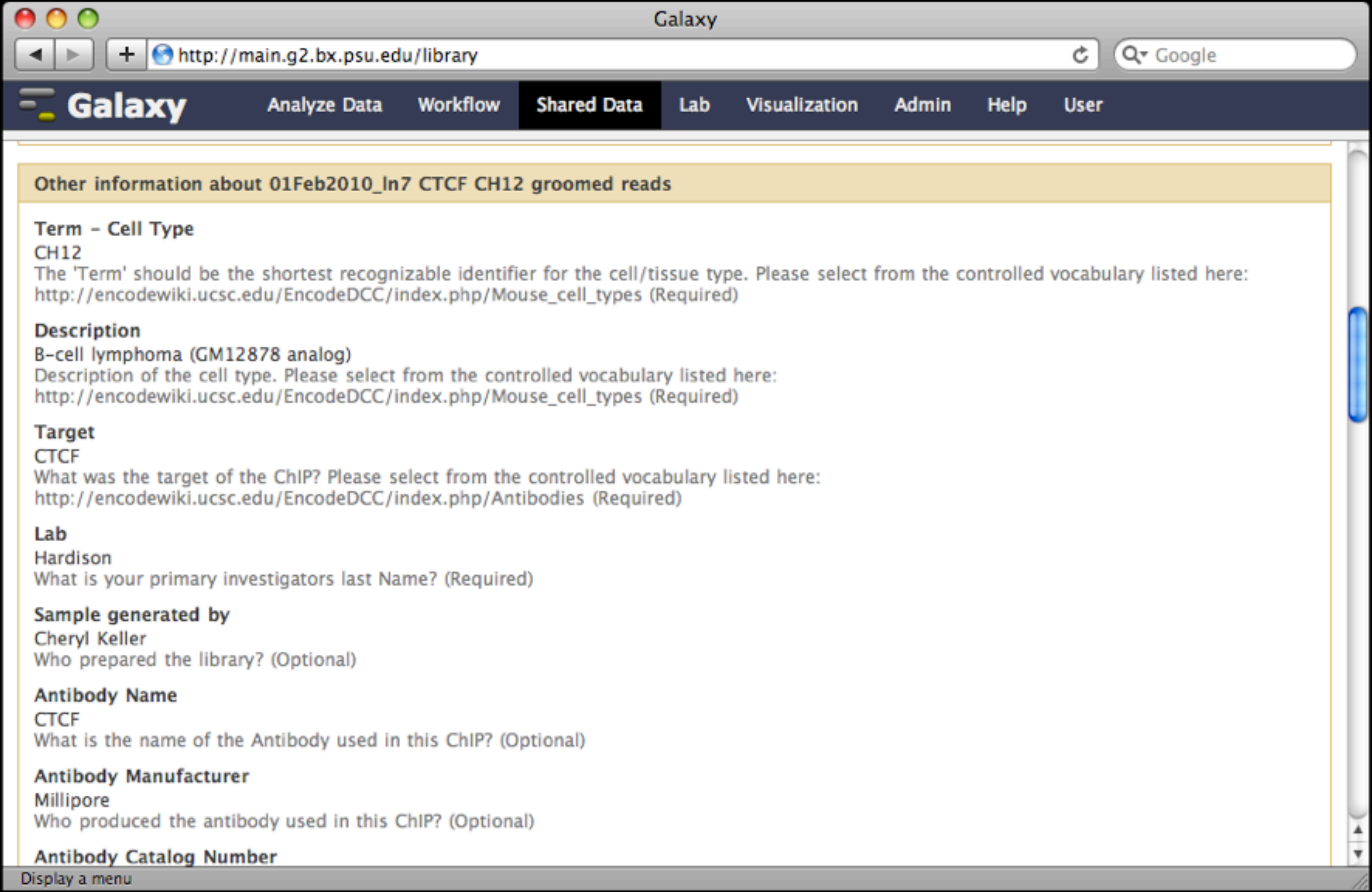
Galaxy build: $Rev 4802:ea7b055efbfa$



- □ G1E Cells
- □ G1E-ER4 Cells
- □ MEL Yale Cells
- □ Enriched
  - □ CTCF ChIP-seq
    - □ CH12 Cells
      - □ Pooled
      - □ Replicate 1

□ 01Feb2010_ln7_CTCF_CH12_groomed reads	None	dan@bx.psu.edu	2010-10-06	2.0 Gb
□ MACS peak calls (broadPeak)	None	dan@bx.psu.edu	2010-10-06	903.0 Kb
□ Mapped Tags (BAM)	None	dan@bx.psu.edu	2010-10-06	493.0 Mb
□ Tag Counts (bigWig)	None	dan@bx.psu.edu	2010-10-06	2.0 Gb

- □ Replicate 2
- □ G1E Cells



Other information about 01Feb2010_ln7 CTCF CH12 groomed reads

Term - Cell Type
CH12
The 'Term' should be the shortest recognizable identifier for the cell/tissue type. Please select from the controlled vocabulary listed here: http://encodewiki.ucsc.edu/EncodeDCC/index.php/Mouse_cell_types (Required)

Description
B-cell lymphoma (GM12878 analog)
Description of the cell type. Please select from the controlled vocabulary listed here: http://encodewiki.ucsc.edu/EncodeDCC/index.php/Mouse_cell_types (Required)

Target
CTCF
What was the target of the ChIP? Please select from the controlled vocabulary listed here: http://encodewiki.ucsc.edu/EncodeDCC/index.php/Antibodies (Required)

Lab
Hardison
What is your primary investigators last Name? (Required)

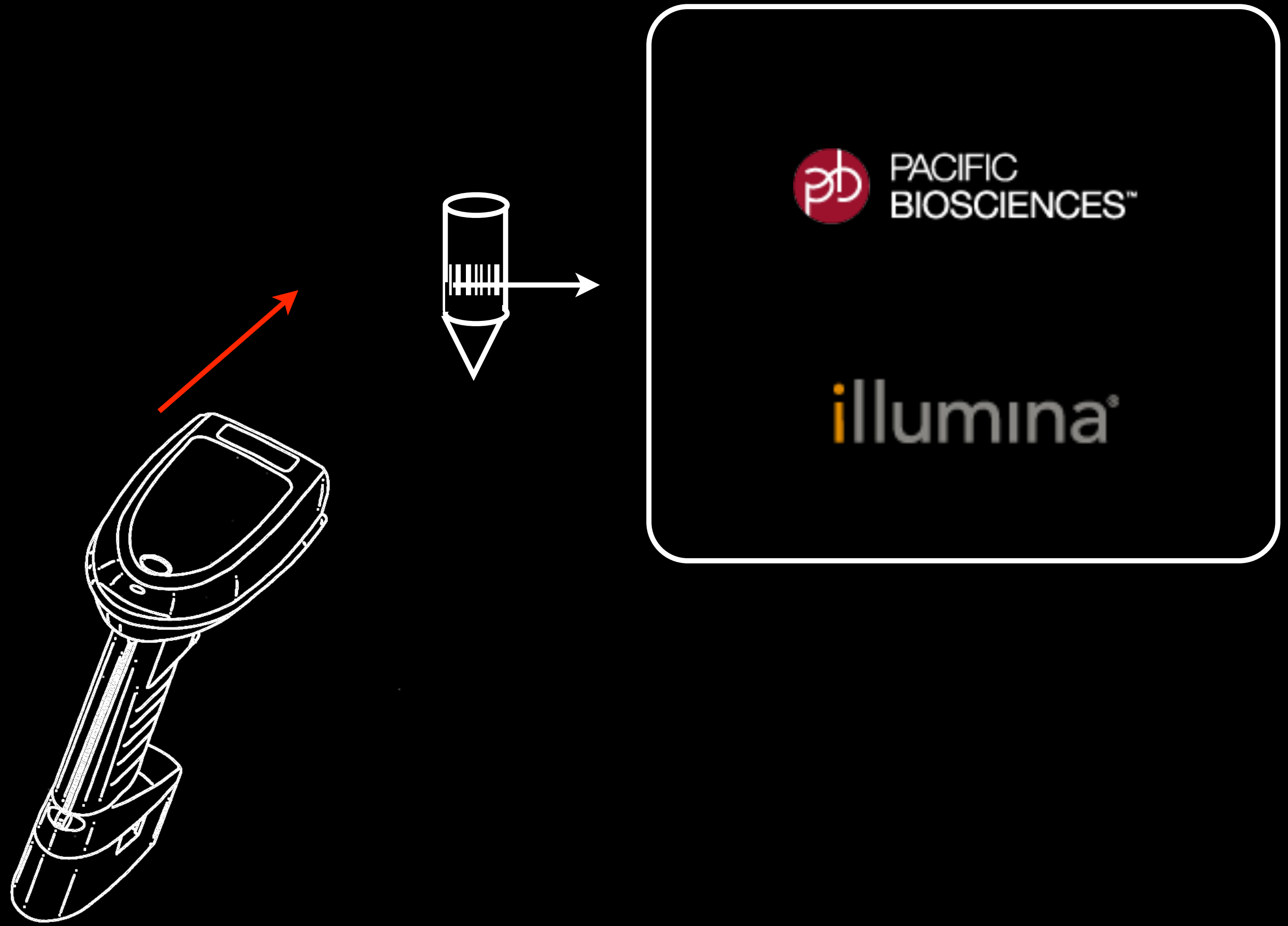
Sample generated by
Cheryl Keller
Who prepared the library? (Optional)

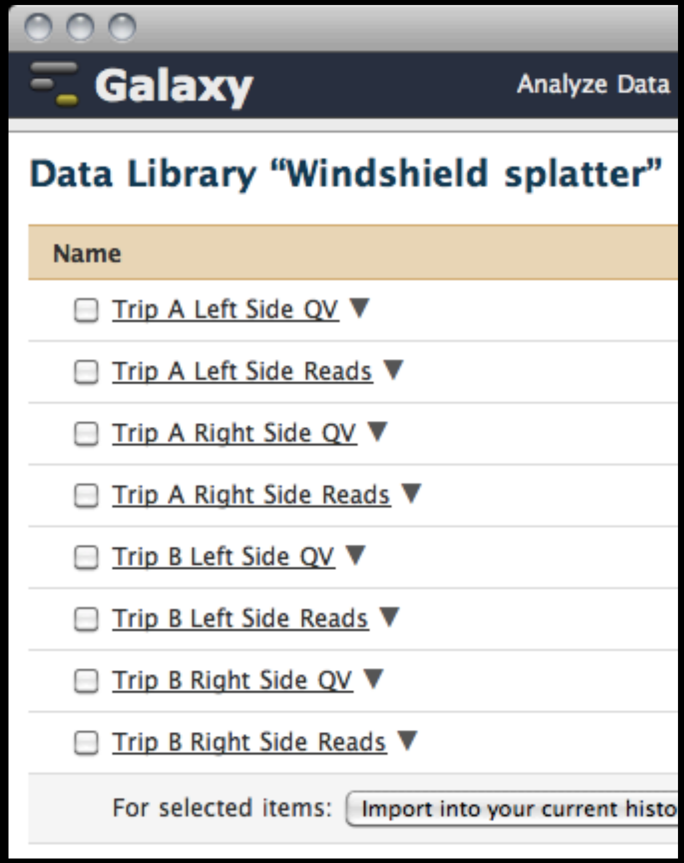
Antibody Name
CTCF
What is the name of the Antibody used in this ChIP? (Optional)

Antibody Manufacturer
Millipore
Who produced the antibody used in this ChIP? (Optional)

Antibody Catalog Number

Sample Tracking



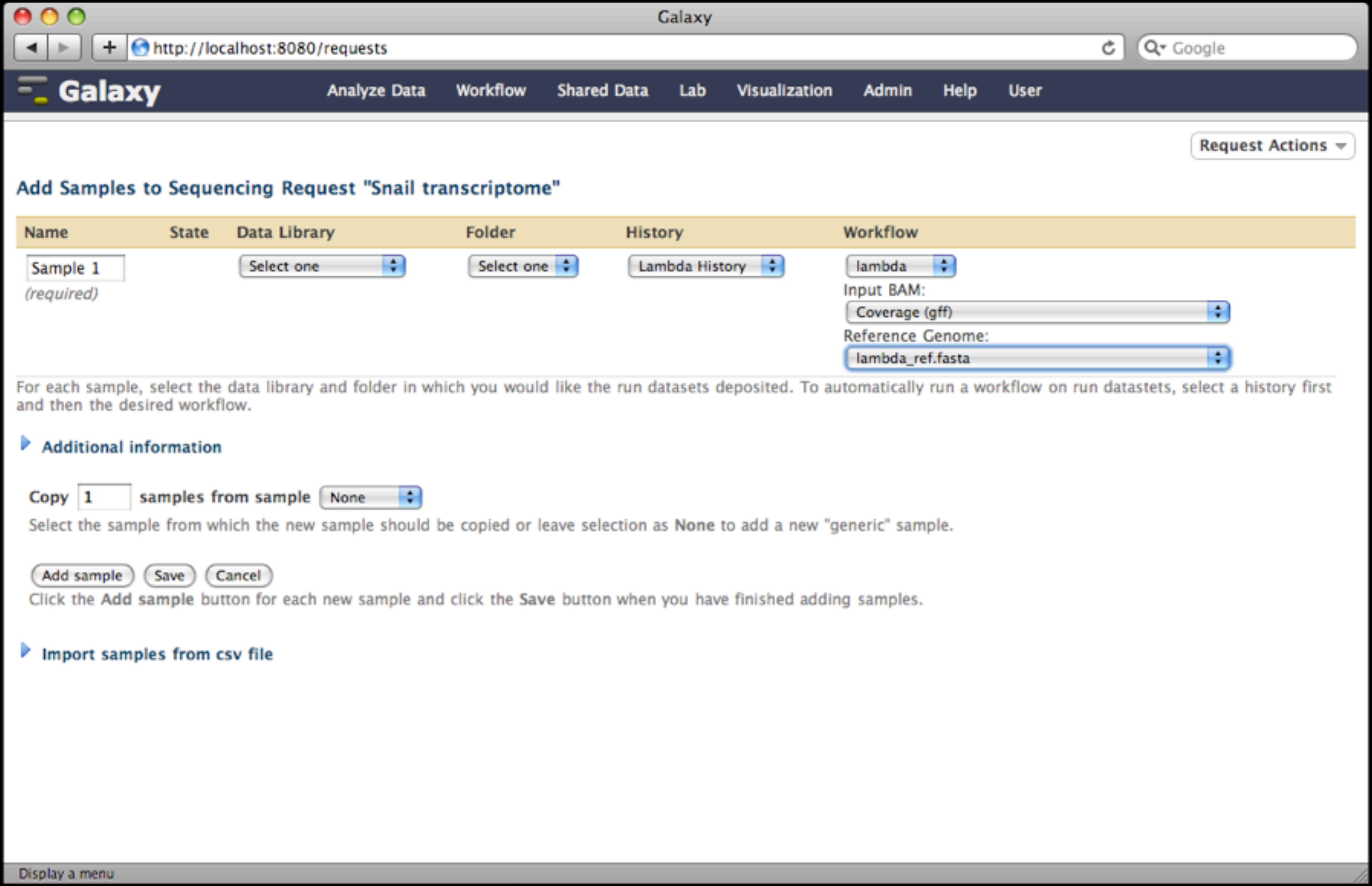


Data Library "Windshield splatter"

Name
□ Trip A Left Side QV ▼
□ Trip A Left Side Reads ▼
□ Trip A Right Side QV ▼
□ Trip A Right Side Reads ▼
□ Trip B Left Side QV ▼
□ Trip B Left Side Reads ▼
□ Trip B Right Side QV ▼
□ Trip B Right Side Reads ▼

For selected items: Import into your current histo

Lab: Sequencing Request Tracking



Add Samples to Sequencing Request "Snail transcriptome"

Name	State	Data Library	Folder	History	Workflow
Sample 1 *(required)*		Select one	Select one	Lambda History	lambda Input BAM: Coverage (gff) Reference Genome: lambda_ref.fasta

For each sample, select the data library and folder in which you would like the run datasets deposited. To automatically run a workflow on run datastets, select a history first and then the desired workflow.

▸ Additional information

Copy 1 samples from sample None

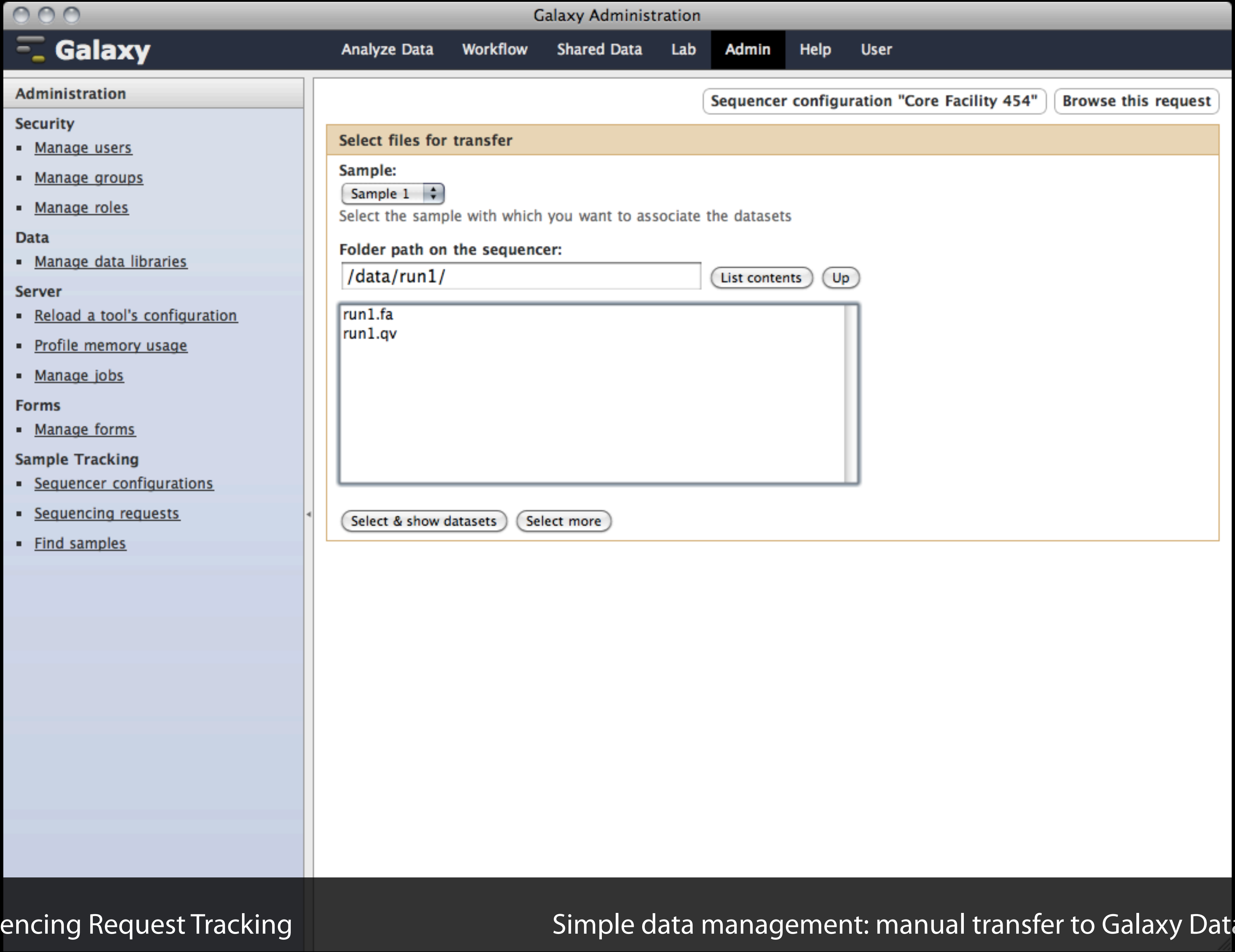
Select the sample from which the new sample should be copied or leave selection as **None** to add a new "generic" sample.

Click the **Add sample** button for each new sample and click the **Save** button when you have finished adding samples.

▸ Import samples from csv file

Lab: Sequencing Request Tracking

Customers add samples, cand define library to deposit in and/or workflows to run



Administration

Security

- Manage users
- Manage groups
- Manage roles

Data

- Manage data libraries

Server

- Reload a tool's configuration
- Profile memory usage
- Manage jobs

Forms

- Manage forms

Sample Tracking

- Sequencer configurations
- Sequencing requests
- Find samples

Select files for transfer

Sample: Sample 1

Select the sample with which you want to associate the datasets

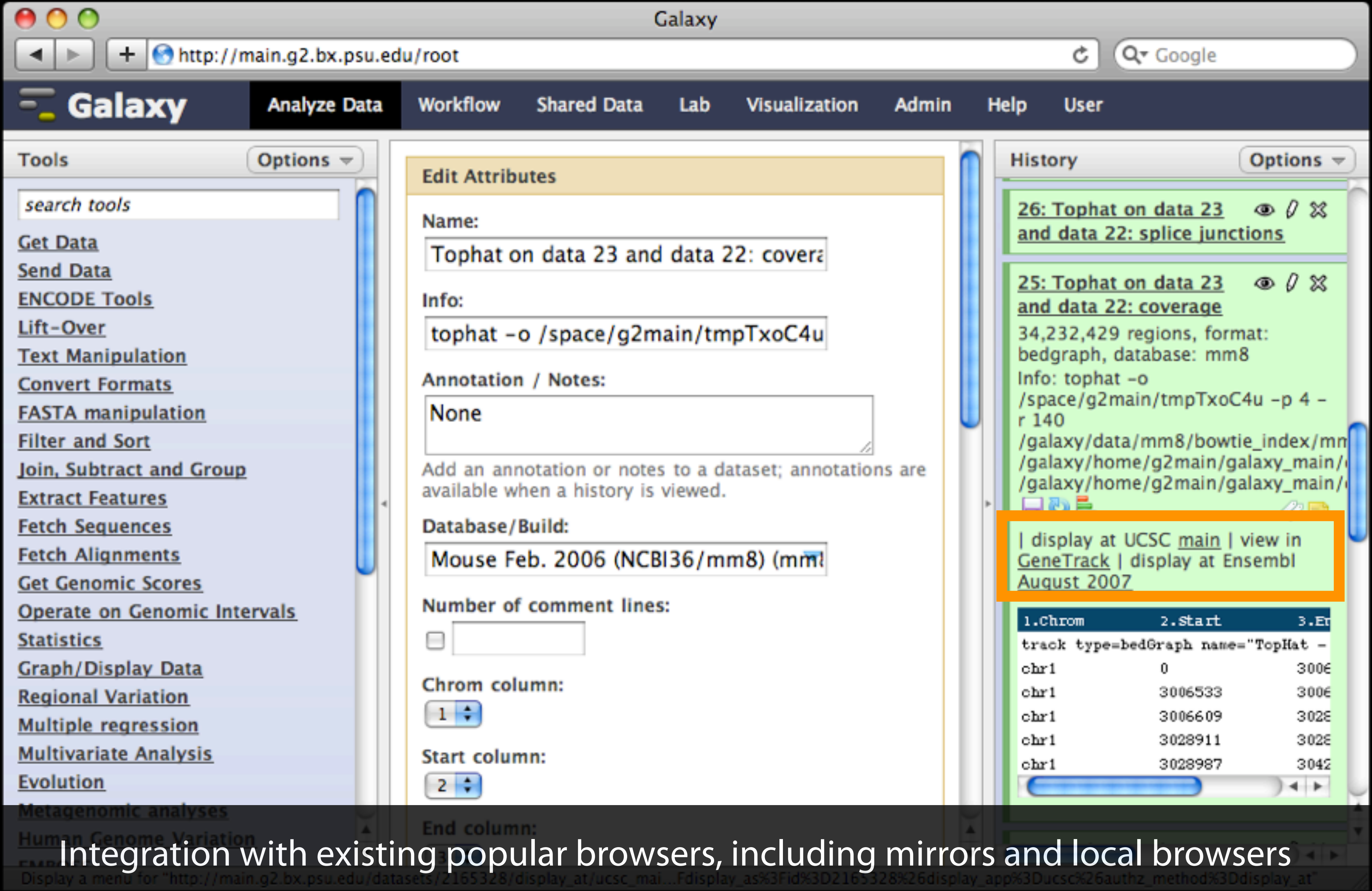
Folder path on the sequencer: /data/run1/

```
run1.fa
run1.qv
```

Lab: Sequencing Request Tracking

Simple data management: manual transfer to Galaxy Data Libraries

Visualization
(beta)



Edit Attributes

Name: Tophat on data 23 and data 22: covera

Info: tophat -o /space/g2main/tmpTxoC4u

Annotation / Notes: None

Add an annotation or notes to a dataset; annotations are available when a history is viewed.

Database/Build: Mouse Feb. 2006 (NCBI36/mm8) (mm8

Number of comment lines: □

Chrom column: 1

Start column: 2

End column:

26: Tophat on data 23 and data 22: splice junctions

25: Tophat on data 23 and data 22: coverage

34,232,429 regions, format: bedgraph, database: mm8

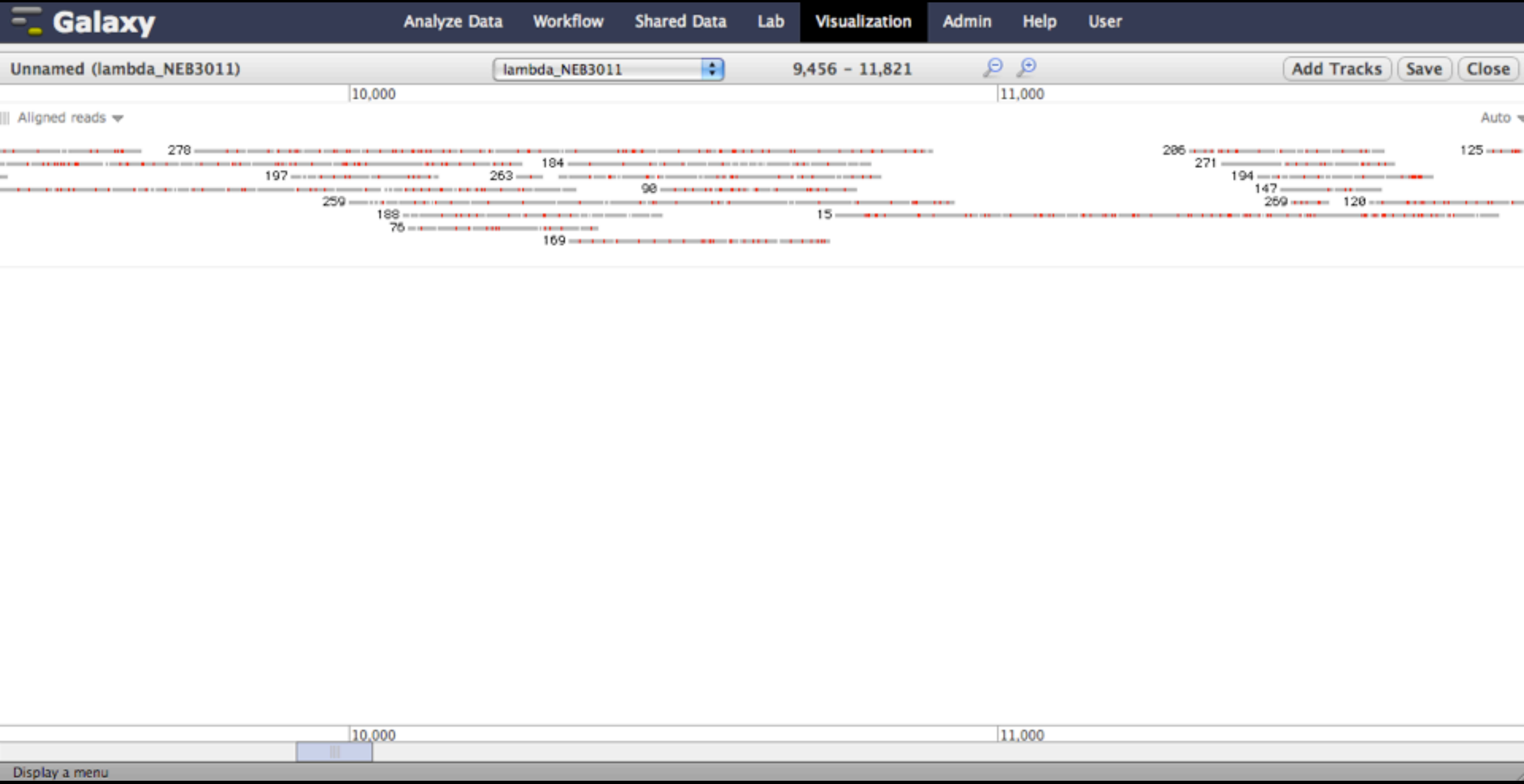
Info: tophat -o /space/g2main/tmpTxoC4u -p 4 -r 140 /galaxy/data/mm8/bowtie_index/mm /galaxy/home/g2main/galaxy_main/ /galaxy/home/g2main/galaxy_main/

| display at UCSC main | view in GeneTrack | display at Ensembl August 2007

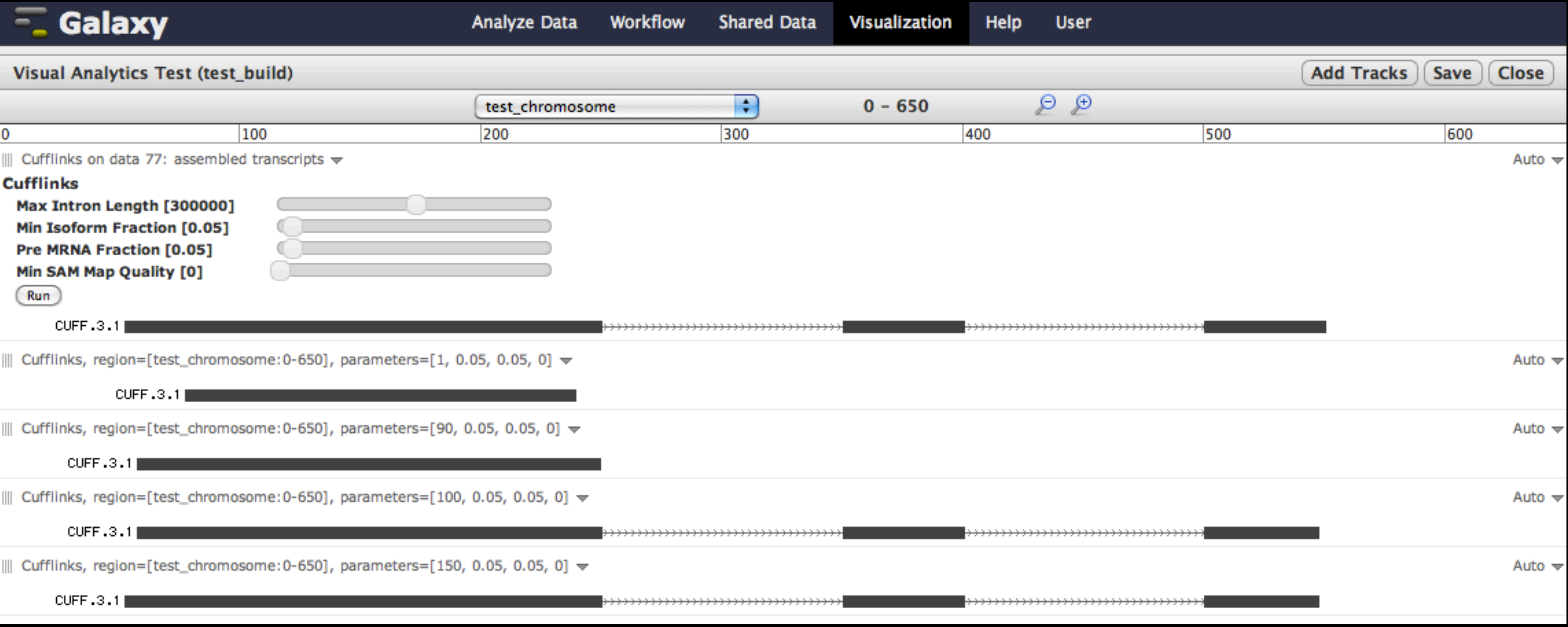
1.Chrom	2.Start	3.En
track type=bedGraph name="TopHat -		
chr1	0	3006
chr1	3006533	3006
chr1	3006609	3028
chr1	3028911	3028
chr1	3028987	3042

Display a menu for "http://main.g2.bx.psu.edu/datasets/2165328/display_at/ucsc_mai...Fdisplay_as%3Fid%3D2165328%26display_app%3Ducsc%26authz_method%3Ddisplay_at"

Integration with existing popular browsers, including mirrors and local browsers



Visualizing aligned reads in trackster



Visual Analytics Test (test_build)

0 - 650

Cufflinks on data 77: assembled transcripts

Cufflinks

Max Intron Length [300000]

Min Isoform Fraction [0.05]

Pre MRNA Fraction [0.05]

Min SAM Map Quality [0]

Cufflinks, region=[test_chromosome:0-650], parameters=[1, 0.05, 0.05, 0]

Cufflinks, region=[test_chromosome:0-650], parameters=[90, 0.05, 0.05, 0]

Cufflinks, region=[test_chromosome:0-650], parameters=[100, 0.05, 0.05, 0]

Cufflinks, region=[test_chromosome:0-650], parameters=[150, 0.05, 0.05, 0]

Visualization integrated with tools: visual analytics in trackster

Try it now:

http://usegalaxy.org

Develop and deploy:

http://getgalaxy.org

The only four things you need to remeber

- **http://usegalaxy.org**
- http://usegalaxy.org/**galaxy101**
- http://usegalaxy.org/**cloud**
- **http://getgalaxy.org**

Galaxy 2011

Community Conference

25-26 May Lunteren, The Netherlands

Help your resource bloom